

Automatic Joint Structured Pruning and Quantization for Efficient Neural Network Training and Compression

Xiaoyi Qu², David Aponte¹, Colby Banbury¹, Daniel P. Robinson², Tianyu Ding¹, Kazuhito Koishida¹, Ilya Zharkov¹, Tianyi Chen^{1*}
 Microsoft¹, Lehigh University²

xiq322@lehigh.edu, Tianyi.Chen@microsoft.com

Abstract

*Structured pruning and quantization are fundamental techniques used to reduce the size of deep neural networks (DNNs), and typically are applied independently. Applying these techniques jointly via co-optimization has the potential to produce smaller, high-quality models. However, existing joint schemes are not widely used because of (1) engineering difficulties (complicated multi-stage processes), (2) **black-box optimization** (extensive hyperparameter tuning to control the overall compression), and (3) insufficient architecture generalization. To address these limitations, we present the framework **GETA**, which automatically and efficiently performs **joint structured pruning and quantization-aware training on any DNN**. **GETA** introduces three key innovations: (i) a **quantization-aware dependency graph (QADG)** that constructs a pruning search space for generic quantization-aware DNN, (ii) a partially projected stochastic gradient method that guarantees **layerwise bit constraints** are satisfied, and (iii) a **new joint learning strategy** that incorporates interpretable relationships between pruning and quantization. We present numerical experiments on both convolutional neural networks and transformer architectures that show that our approach achieves competitive (often superior) performance compared to existing joint pruning and quantization methods. Source code is available at <https://github.com/microsoft/GETA>.*

1. Introduction

Deep neural networks (DNNs) have been widely used in various applications [26, 34, 36, 56]. However, their increasing size has raised several concerns. One major challenge is the substantial storage space required to hold these models, which can be **impractical** for everyday devices such as standard PCs and even more so for resource-constrained edge devices [50]. Furthermore, as model sizes increase, inference cost often lengthens, leading to delays that can

be **frustrating** for users who expect quick responses. Therefore, it is of practical interest and importance to compress the model while maintaining performance similar to the full model. To address the above concerns, various model compression techniques have been studied in recent years [14].

Pruning and quantization are two fundamental techniques that are widely deployed, each following different methodologies. Structured pruning is perhaps the most popular pruning scheme, which aims to remove redundant structures in DNNs while preserving performance [10, 18]. Quantization reduces the bit width of the data flowing through a DNN [14]. In practice, structured pruning is typically applied first to identify a high-performing subnetwork, which is then quantized to further reduce its size and enhance its processing speed on specified hardware [23, 42]. However, treating pruning and quantization separately has limitations. For example, more heavily structurally pruned models are typically more sensitive to quantization and thus require higher bit widths. Thus, joint structured pruning and quantization becomes an important topic.

1.1. Challenges

Many studies [3, 23, 29, 37, 42, 47, 52, 55, 58, 61, 64] have combined pruning and quantization to obtain high compression ratios. However, these joint methods are not commonly used in practice due to one or more of the following reasons: engineering difficulty, black-box optimization, and insufficient architecture generalization, which we now discuss.

Engineering Difficulties. First, many joint pruning and quantization methods follow a two-stage process. For example, [52, 55, 58, 61] first determine the configurations (pruning ratio and bit width) for each layer of the network, and then train the pruned and quantized model. They require separate compression and retraining stages since the two stages may be incompatible with each other. Thus, two-stage pipelines increase the execution time, especially for large datasets (e.g., ImageNet). For these reasons, a one-shot (all-at-once) framework is preferred. Second, while recent automated structured pruning frameworks propose dependency graphs to support generic architectures [11, 18],

*Corresponding Author.

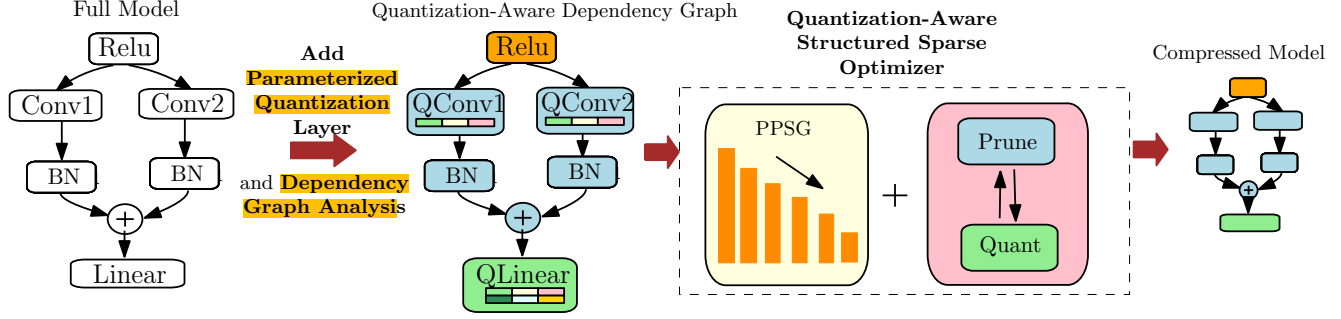


Figure 1. **GETA** framework pipeline. Nodes Conv1 and Conv2 represent two convolutional layers, node BN represents batch normalization, and the “+” represents summation. For details on the remainder of the figures, see Sec. 3–Sec. 5.

integrating quantization introduces new challenges. The addition of attached and inserted branches in the trace graph, which are not accounted for in existing dependency graph analyses, breaks the supports for any architecture.

Black-box Optimization Process. A significant portion of existing methods lack explicit control over sparsity and bit width during the optimization process. This limitation arises from the multi-objective nature of joint compression problems, which require balancing conflicting goals: maintaining model performance while maximizing sparsity and minimizing bit width. Approaches such as DJPQ [58], BB [55], and Clip-Q [52] often tackle this challenge by introducing regularization coefficients to reconcile conflicting objectives. However, a significant drawback is their inability to predict the final compression ratio of the model prior to executing the entire optimization process. Consequently, users typically require extensive hyper-parameter tuning efforts, limiting flexibility and productivity in practice.

Insufficient Architecture Generalization. The existing work [3, 37, 47, 52, 55, 58, 61] on joint structured pruning and quantization focuses primarily on the convolutional neural network (CNN), and cannot be applied to architectures such as transformers. For instance, both DJPQ [58] and BB [55] applies per-channel pruning to each layer, which will not work for multi-head attention in transformers because it ignores the dependencies between different attention heads.

Table 1. **GETA** versus representative joint pruning and quantization methods in terms of (i) whether the method supports structured pruning, (ii) whether it is a one-shot approach, (iii) whether it is a white-box approach, and (iv) whether it can be used on a variety of network architectures and tasks (*i.e.*, generalization). Methods not listed lack one or more of these properties.

	GETA	BB	DJPQ	QST	Clip-Q	ANNC
Structured Prune [†]	✓	✓	✓	✓	✓	✓
One-shot [†]	✓	✗	✗	✓	✓	✗
White-box Optimization	✓	✗	✗	✓	✗	✓
Generalization	✓	✗	✗	✗	✗	✗

[†] Categorized into engineering difficulties.

1.2. Our Contributions

To tackle the above challenges, we propose **GETA**, a General and Efficient Training framework that Automates joint structured pruning and quantization aware training.

By streamlining the workflow, **GETA** significantly reduces the engineering burden and minimizes user intervention (See **Framework Usage**).

As shown in Fig. 1, **GETA** begins by incorporating the parameterized quantization layer [53] into the full model, which allows for layerwise bit widths to be learned during training (see Sec. 3). Next, the framework proposes a quantization-aware dependency graph (QADG) (see Sec. 4) to address previously unconsidered graph transformations introduced by parameterized quantization layers, ensuring support for any architecture. To train the neural network using the quantization-aware dependency graph, we employ a quantization-aware structured sparse optimizer (see Sec. 5) to determine the optimal tradeoff between the pruning ratio and bit width for each layer. Our main contributions are summarized as follows.

- **Quantization-Aware Dependency Graph (QADG).** We propose the quantization-aware dependency graph (QADG) to support joint structured pruning and quantization applied to any quantization-aware deep neural network (QADNN). By eliminating the need to handle each architecture individually, QADG significantly reduces the model-specific engineering workloads.
- **Quantization-Aware Structured Sparse Optimizer (QASSO).** We propose a quantization-aware structured sparse optimizer, to provide reliable joint structured pruning and mixed precision quantization-aware training. To the best of our knowledge, QASSO is the first white-box joint optimizer that explicitly controls the sparsity ratio and bit width. Particularly, QASSO employs a partial projected stochastic gradient (PPSG) method that progres-

Framework Usage

```

1 import GETA
2 # General DNN model
3 geta = GETA(model)
4 optimizer = geta.qasso()
5 # Train as normal
6 optimizer.step()
7 # Quantized Pruned DNN
8 geta.construct_subnet()

```

sively converges towards the bit width budget for training stability. Moreover, a joint learning strategy is introduced to address the conflicts between pruning and quantization for performance preservation.

- **Numerical Verification.** We test **GETA** on a wide range of neural networks including ResNet, VGG, BERT, Phi2, and ViT, among others. The results indicate that **GETA** achieves competitive (often superior) performance compared to state-of-the-art joint pruning and quantization methods in terms of performance and bit operations.

2. Related Work

Structured Pruning. Structured pruning aims to remove redundant structures to reduce the size of DNNs. The identification of redundancies can be performed based on different criteria such as sparsity [5–7, 10, 20, 22, 31, 38, 44, 59, 62, 68], Bayesian pruning [55, 67]. Previous methods typically use a complicated, time-costly process that requires domain knowledge to train the DNN. Another challenge is to define a pruning search space procedure that can be generalized to various DNNs. Recent frameworks, such as OTO [9, 11, 12] and DepGraph [18], have automated the construction of this search space using dependency graphs. However, these methods are not suitable for QADNNs due to prevalent issues such as weight-sharing and shape ambiguous operators.¹ This limitation highlights the ongoing challenge of automating structured pruning for any QADNN.

Quantization-Aware Training (QAT). The standard approach to QAT is applying a uniform bit width across all layers. However, [16, 28] empirically show that different layers in DNNs exhibit different sensitivities to quantization, suggesting that mixed-precision quantization may be a better approach for reducing performance loss. Several strategies including parameterized quantizers [53], heuristic approaches [47], reinforcement learning [3, 17], multi-objective Bayesian optimization [45], and Hessian information guided methods [15, 16, 63] have been proposed to determine the optimal bit width for each layer.

Joint Pruning and Quantization. The challenge of using a joint approach lies in determining an optimal trade-off between the pruning ratio and quantization levels for the model. Two primary strategies have been explored to address this challenge. The first strategy is to efficiently search the joint parameter space with prior work considering heuristics [47], reinforcement learning [3], and Bayesian optimization [52]. The second strategy focuses on gradient-based optimization techniques. [61, 64] formulates a constrained optimization problem and solves it via alternating direction method of multipliers. However, both approaches fail to support structured pruning. [40, 55, 58]

introduces regularization coefficients to balance conflicting objectives, preserving model performance while reducing computational burden, which require significant hyperparameter tuning. [29] proposes a one-shot framework for the joint compression of DNNs, but it does not support structured pruning. Unlike these approaches, **GETA** is a one-shot, white-box framework that supports joint structured pruning and quantization by gradient-based optimization.

3. Quantization with Learnable Parameters

Instead of freezing the bit width in standard QAT approach, we introduce quantization parameters q_m , t , and d to learn the bit width of each layer [54]. In particular, q_m represents the maximum value to be mapped and t is the exponent controlling the shape of the mapping and d , known as quantization step size, characterizes the interval between adjacent quantization levels. For each quantization operation, we first quantize the input tensor x as \tilde{x} by applying a nonlinear function [58]

$$\tilde{x} = \text{sgn}(x) \cdot \begin{cases} |x|^t, & |x| \leq q_m, \\ (q_m)^t, & |x| > q_m. \end{cases} \quad (1)$$

After applying the nonlinear mapping, we perform symmetric uniform quantization on \tilde{x} , resulting in the mapping

$$x^Q = d \lfloor \tilde{x}/d \rfloor, \quad (2)$$

where $\lfloor \cdot \rfloor$ represents rounding to the nearest integer. The associated bit width b is computed as

$$b = \log_2 \left(\frac{(q_m)^t}{d} + 1 \right) + 1. \quad (3)$$

To optimize quantization variables d , t , and q_m , we compute their gradients using the straight-through estimator [53]. For details of the gradient of the quantization mapping with respect to d , t , and q_m , one can find them at Appendix A.

Remark. The computation involving x in this section represents element-wise operations.

4. Quantization-Aware Dependency Graph

To automate joint structured pruning and quantization-aware training, we first establish a pruning search space. This space is defined as the set of minimally removable structures within the target DNN, ensuring that the remaining sub-network remains functional post-removal. However, establishing this search space automatically is challenging due to the complex topology of DNNs and the diverse roles of operators. Recent advancements in dependency graph [11, 18] address some of these challenges, but existing approaches remain insufficient for QADNN.

¹Shape ambiguous operators are operators (e.g., *reshape* and *view* in PyTorch) that transform input tensors into outputs of varying dimensions.

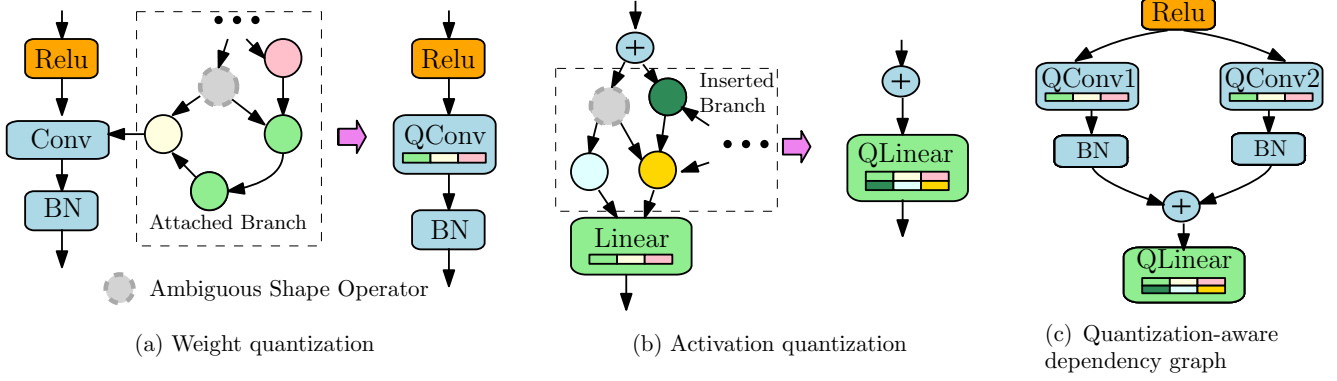


Figure 2. Figure 2(a) and 2(b) illustrate the Quantization-Aware dependency graph analysis for weight quantization and activation quantization, respectively. Figure 2(c) presents a dependency graph after QADG analysis. Concrete examples are provided in Appendix F.

To automate the construction of the pruning search space for QADNN, we construct a Quantization-Aware Dependency Graph (QADG). QADG efficiently captures pruning dependencies across both weight and activation quantization. Challenges arise due to the numerous parameterized layers introduced during layer conversion, which include weight-sharing and shape-ambiguous layers that previous algorithms do not account for. Weight and activation quantization-aware layers exhibit distinct structural patterns. As shown in Fig. 2(a), weight quantization introduces a prominent attached branch connected to the target layer. In contrast, activation quantization inserts a set of layers between the activation layer and its subsequent layer, referred to as the inserted branch.

Algorithm 1 Constructing a Quantization-Aware Dependency Graph

- 1: **Input:** Trace graph $(\mathcal{V}, \mathcal{E})$ of QADNN.
- 2: **Initialize:** $\mathcal{V}_{\text{root}}^{\text{weight}} = \emptyset$, $\mathcal{V}_{\text{root}}^{\text{act}} = \emptyset$, and $\mathcal{V}_{\text{end}}^{\text{act}} = \emptyset$.
- 3: Traverse $(\mathcal{V}, \mathcal{E})$ and add the root vertex of each attached branch to the set $\mathcal{V}_{\text{root}}^{\text{weight}}$.
- 4: **for each** $v \in \mathcal{V}_{\text{root}}^{\text{weight}}$ **do**
- 5: Find attached branch associated with root vertex v .
- 6: Merge vertices in attached branch as vertex \tilde{v} .
- 7: Replace v with \tilde{v} .
- 8: **end for**
- 9: Traverse $(\mathcal{V}, \mathcal{E})$ and add the root vertex and end vertex of each inserted branch to $\mathcal{V}_{\text{root}}^{\text{act}}$ and $\mathcal{V}_{\text{end}}^{\text{act}}$, respectively.
- 10: **for each pair** $(v_{\text{root}}, v_{\text{end}}) \in \mathcal{V}_{\text{root}}^{\text{act}} \times \mathcal{V}_{\text{end}}^{\text{act}}$ **do**
- 11: Merge vertices between v_{root} and v_{end} as vertex \tilde{v} .
- 12: Replace v_{end} with \tilde{v} .
- 13: Add an edge from v_{root} to \tilde{v} into \mathcal{E} .
- 14: **end for**
- 15: Conduct dependency graph analysis in [11].
- 16: **Output:** the QADG obtained from Line 15.

Quantization-Aware Dependency Graph Analysis. To tackle these challenges, as stated in Algorithm 1, we propose QADG analysis. At Line 3, we first traverse the trace graph $(\mathcal{V}, \mathcal{E})$ via depth-first search to identify the set of root vertices, $\mathcal{V}_{\text{root}}^{\text{weight}}$, for weight quantization. An example of a root vertex is *Conv* in Fig. 2(a). We then identify attached branches, merge them as new vertices, and replace the root vertices with these new structures, as specified at Line 4-Line 8. For activation quantization, we first locate the root and end vertices, such as *Relu* and *QLinear* in Fig. 2(a). Next, we identify the inserted branches, merge them as new vertices, and replace the end vertices with these new structures. To preserve the connectivity of QADNN, we reconnect the root vertices with the newly formed end vertices in Line 13. Through this optimization, we consolidate complex attached and inserted branches into single entities, allowing us to de-duplicate shared weights and eliminate shape-ambiguous vertices. Subsequently, we apply the dependency graph analysis from [11] to derive the QADG, which facilitates the construction of the pruning search space over the QADNN, enabling joint structured pruning and quantization-aware training.

5. QASSO

After obtaining a QADG using Algorithm 1, we obtain the pruning search space of the QADNN, *i.e.*, the parameter groups \mathcal{G} . Each $g \in \mathcal{G}$ represents the collection of trainable variables in one minimally removal structure. We then apply the QASSO optimizer (see Algorithm 2) to the problem

$$\underset{\substack{x \in \mathbb{R}^n \\ (d, q_m, t) \in \mathbb{R}^{|\mathcal{L}|} \times \mathbb{R}^{|\mathcal{L}|} \times \mathbb{R}^{|\mathcal{L}|}}}{\text{minimize}} \quad f(x, d, q_m, t) \quad (4a)$$

$$\text{s.t.} \quad \text{Card}\{g \in \mathcal{G} \mid [x]_g = 0\} = K, \quad (4b)$$

$$b_i \in [b_l, b_u], i \in \mathcal{L}, \quad (4c)$$

where K represents the target sparsity ratio, $[b_l, b_u]$ specifies the target bit width range, and \mathcal{L} denotes the index set

of layers that have parameterized quantization layers added, and $|\mathcal{L}|$ represents the cardinality of set \mathcal{L} , and bit width b_i is computed using formula Eq. (3) given in Sec. 3.

Algorithm 2 QASSO

- 1: **Inputs:** Initial weight parameters x and quantization parameters (d, q_m, t) , learning rate schedule $\{\alpha_l\}$, number of warm-up steps K_w , bit width range $[b_l, b_u]$ with $b_u \geq b_l + 1$, number of projection periods $B \in [1, b_u - b_l]$, bit width reduction factor $b_r \in [1, (b_u - b_l)/B]$, number of projection steps K_b , number of pruning steps K_p , and number of pruning periods P .
 - 2: Perform K_w SGD steps on (4a) to update (x, d, q_m, t) .
 - 3: **for** each projection period $0, 1, \dots, B - 1$ **do**
 - 4: $b_u \leftarrow b_u - b_r$.
 - 5: **for** $k = 0, 1, \dots, K_b - 1$ **do**
 - 6: Update x using one step of SGD on (4a).
 - 7: Update (d, q_m, t) using Algorithm 3.
 - 8: **end for**
 - 9: **end for**
 - 10: **for** each pruning period $0, 1, \dots, P - 1$ **do**
 - 11: Compute saliency score [12] using x .
 - 12: Compute the set of important groups \mathcal{G}_I and set of redundant groups \mathcal{G}_R using the saliency score.
 - 13: **for** $k = 0, 1, \dots, K_p - 1$ **do**
 - 14: Update (t, q_m) using one step of SGD on (4a).
 - 15: Stochastic gradient $\hat{\nabla}_x f \approx \nabla_x f(x, d, q_m, t)$.
 - 16: Compute γ using Eq. (13).
 - 17: Update d with Eq. (14).
 - 18: Compute x^Q from Eq. (2).
 - 19: For currently scheduled learning rate α , set

$$[x]_{\mathcal{G}_I} \leftarrow [x]_{\mathcal{G}_I} - \alpha[\hat{\nabla}_x f]_{\mathcal{G}_I} \quad \text{and} \quad (5)$$

$$[x]_{\mathcal{G}_R} \leftarrow [x]_{\mathcal{G}_R} - \alpha[\hat{\nabla}_x f]_{\mathcal{G}_R} - \gamma[x^Q]_{\mathcal{G}_R}. \quad (6)$$
 - 20: **end for**
 - 21: **end for**
 - 22: Fixing the quantization parameters, say (d^*, q_m^*, t^*) , computed above, train Eq. (4a) over the weight parameters in the set of important groups \mathcal{G}_I to obtain x^* .
 - 23: **Outputs:** Parameters (x^*, d^*, q_m^*, t^*) .
-

Overview of QASSO. Our framework QASSO (see Algorithm 2) aims to compress the size of the DNN while preserving full model performance by removing redundant structures, determining the optimal bit width for each layer that has a parameterized quantization layer added, and recovering knowledge lost during pruning and quantization phases. This is accomplished through a sequential four-stage optimization process: warm-up stage, projection stage, joint stage, and a cool-down stage. The warm-up stage consists of optimizing over all trainable variables using the stochastic gradient (SGD) method or any of its vari-

ants at Line 2, which allows us to achieve a better initialization for improved performance. Next, we step into the projection stage (see Line 3-9), where we progressively reduce the bit width range until the bit width constraint Eq. (4c) is satisfied. This progressive technique enables us to transfer information lost in the low bit precision representation back to the current model. We then proceed to the joint stage (see Line 10-21), where we progressively forget the quantized information (see Eq. (6)) within the redundant groups until the constraint Eq. (4b) is satisfied. In addition, the bit width selected depends on the amount of information removed within each layer at each step. Specifically, when a significant amount of information is removed, we will consider employing a high bit width for quantization. Once we complete pruning and determine the bit width for each layer, we train the pruned and quantized model until convergence, referred to as the cool-down stage. The projection stage and joint stage are two essential components in our approach and we will discuss them in the next two subsections.

5.1. Projection Stage

During the projection stage, we aim to compute a feasible bit width. To do so, we consider the problem

$$\min_{\substack{x \in \mathbb{R}^n \\ (d, q_m, t) \in \mathbb{R}^{|\mathcal{L}|} \times \mathbb{R}^{|\mathcal{L}|} \times \mathbb{R}^{|\mathcal{L}|}}} f(x, d, q_m, t) \quad (7a)$$

$$\text{s.t.} \quad b_i \in [b_l, b_u], i \in \mathcal{L}. \quad (7b)$$

Related Approaches and Limitations. In numerical optimization, projection methods and penalty methods are two of the most common approaches for training DNNs with explicit constraints. However, both approaches are inappropriate for our problem setting Eq. (7). On one hand, the projection method is effective when the projection operator has a closed-form solution, while the projection operator associated with Eq. (7b) lacks this property. On the other hand, penalty methods (e.g., [4, 46]) consider a sequence of sub-problems that relax the constraint by penalizing its violation in the objective function. Its effectiveness is highly dependent on an appropriate selection of the penalty parameter, which often necessitates hyperparameter tuning.

Algorithm 3 Partial Projected Stochastic Gradient.

- 1: **Inputs:** Variables d, q_m, t , and bit width range $[b_l, b_u]$.
 - 2: Update variables d, q_m, t using SGD or its variants.
 - 3: Determine the range $[d_{\min}, d_{\max}]$ of d using (q_m, t) and formula Eq. (3).
 - 4: Project d onto $[d_{\min}, d_{\max}]$.
 - 5: **Outputs:** d, q_m, t .
-

Given the above discussion, we propose a variant of a projected stochastic gradient method called partial projected stochastic gradient (PPSG) (see Algorithm 3). In this

approach, the projection is applied only to the variable d . Alternatively, one could apply the projection operation to either q_m or t , but our numerical testing shows this often leads to training instability (gradients explode or vanish).

5.2. Joint Stage

During the joint stage, we aim to identify redundant groups of \mathcal{G} , to forget the information within the redundant groups and transfer to the important groups being aware of the quantization parameters, and to determine the layerwise bit widths in terms of the information removed at each layer.

We first partition our parameter group \mathcal{G} into a set of important groups \mathcal{G}_I and a set of redundant groups \mathcal{G}_R based on saliency scores detailed in [12] at Line 12. For variables in \mathcal{G}_I , we proceed with vanilla stochastic gradient or its variants at Eq. (5). For variables in \mathcal{G}_R , we progressively project them to zero by forgetting redundant information at Eq. (6). Due to the addition of parameterized quantization layers to the original model, weight parameters x are converted to its quantized counterpart, denoted as x^Q . This observation underscores the necessity to forget the quantized information $[x^Q]_{\mathcal{G}_R}$ instead of the original information $[x]_{\mathcal{G}_R}$. Additionally, it is essential to develop a new update rule for the forget rate γ that is aware of quantization parameters to better maintain and transfer the knowledge.

For ease of notation, we denote the stochastic gradient of function $f(x, d, q_m, t)$ with respect to x as $\hat{\nabla}_x f$. Consequently, the search direction $s(x)$ for updating x is

$$s(x) = \begin{cases} -\alpha[\hat{\nabla}_x f]_g, & g \in \mathcal{G}_I, \\ -\alpha[\hat{\nabla}_x f]_g - \gamma[x^Q]_g, & g \in \mathcal{G}_R. \end{cases} \quad (8)$$

The quantized value x^Q in Eq. (8) can be rewritten as

$$x^Q = \text{sgn}(x) \cdot \text{clip}_{q_m}^t(|x|) + d \cdot \text{sgn}(x) \cdot R(x), \quad (9)$$

where the clipped value can be written as

$$\text{clip}_{q_m}^t(|x|) = \begin{cases} |x|^t, & |x| \leq q_m, \\ (q_m)^t, & |x| > q_m, \end{cases} \quad (10)$$

and the residual value is given by

$$R(x) = \begin{cases} \lfloor \frac{|x|^t}{d} \rfloor - \frac{|x|^t}{d}, & |x| \leq q_m, \\ \lfloor \frac{(q_m)^t}{d} \rfloor - \frac{(q_m)^t}{d}, & |x| > q_m. \end{cases} \quad (11)$$

We denote the angle between $-\hat{\nabla}_x f|_g$ and $-\text{sgn}(x) \cdot \text{clip}_{q_m}^t(|x|)|_g$ as θ_γ and the angle between $-\hat{\nabla}_x f|_g$ and $-\text{sgn}(x) \cdot d \cdot R(x)|_g$ as θ_d . The *clip* represents the mean of the clipped value within the redundant group \mathcal{G}_R , i.e.,

$$\text{clip} = \text{mean}([\text{clip}_{q_m}^t(|x|)]_{\mathcal{G}_R}). \quad (12)$$

With the above notations, the forget rate γ selection rule is expressed, for pre-specified small ϵ and $\eta \in (0, 1)$, as

$$\gamma = \begin{cases} 0, & \text{clip} \leq \epsilon, \\ 1 - \frac{K_p - k - 1}{K_p - k}, & \cos(\theta_\gamma) \geq 0, \text{clip} > \epsilon, \\ -\frac{(1-\eta)\alpha\|\hat{\nabla}_x f|_g\|}{\cos(\theta_\gamma)\|\text{sgn}(x) \cdot \text{clip}_{q_m}^t(|x|)|_g\|}, & \cos(\theta_\gamma) < 0, \text{clip} > \epsilon. \end{cases} \quad (13)$$

The quantization step size d selection rule is, for $\xi \in (0, 1)$,

$$d = \begin{cases} \frac{(q_m)^t}{2^{b_l-1}-1}, & \cos(\theta_d) \geq 0, \\ -\frac{\xi\eta\alpha\|\hat{\nabla}_x f|_g\|}{\gamma\cos(\theta_d)\|\text{sgn}(x) \cdot R(x)|_g\|}, & \cos(\theta_d) < 0. \end{cases} \quad (14)$$

Interpretation of Update Rules for γ and d . At a high level, the update rule for the forget rate and quantization step size ensures that the search direction in Eq. (8) is a descent direction for the objective function f , as stated in Proposition 5.1. Consequently, forgetting knowledge stored in the redundant groups for pruning and quantizing the variables jointly in this manner can make progress towards convergence. Therefore, the conflict between pruning and quantization is largely resolved via our design.

Remarks. When the mean of the clipped values within the redundant group \mathcal{G}_R is relatively small, we reasonably infer that little knowledge is retrieved in the redundant group. Therefore, we set the forget rate to 0 and directly project all parameters in the redundant group \mathcal{G}_R to zero. Otherwise, our forget rate rule is divided based on the angle between the gradient and clipped values. When $\cos(\theta_\gamma) \geq 0$, any positive values can be chosen where we select it as a uniform forgetting rate within K_p steps. The quantization step size d is divided into two cases in terms of the angle between the gradient and the residual values. When $\cos(\theta_d) \geq 0$, d can be selected as any positive values. In this scenario, we consider a low bit width for quantization and specifically, d is selected such that the computed bit width is equal to b_l , the min of the bit width range $[b_l, b_u]$. For details of the joint stage implementation, one can refer to Appendix C.

Proposition 5.1. *Let $\hat{\nabla}_x f$ be the full gradient of function $f(x, d, q_m, t)$ with respect to x . With forget rate γ selection rule Eq. (13) and quantization step size d selection rule Eq. (14), the search direction $s(x)$ is a descent direction for the function f with respect to x at x .*

Proof. See Appendix B □

6. Numerical Experiments

In this section, we present numerical experiments to demonstrate the effectiveness of our approach, accompanied by ablation studies provided in Appendix E to assess the contribution of each stage to the success of GETA. ²

²Experiment setup details are provided in Appendix D.

DNN Architectures and Datasets. The experiments are performed across a wide range of popular CNN architectures, such as VGG7 [33], ResNet20, ResNet50 and ResNet56 [26], and transformer architectures, such as Bert [34], varying vision transformers [1] and Large Language Models (LLMs) such as Phi2-2.7B [32]. The selected datasets include the benchmark CIFAR10 [35], ImageNet2012 [13], Squad [34], and common-sense benchmarks in LM-Evaluation-Harness [21].

Comparing Methods. To validate the effectiveness and superiority of our framework, we consider the following methods for comparisons: ANNC [61], QST-B [47], DJPQ [58] and its variants, BB [55], Clip-Q [52], OBC [19], and a standard first-prune-then-quantize method. All these methods consider both pruning and quantization. Moreover, they use the same strategy that first conducts a search based on the pretrained model and then fine-tunes the resulting model with the configurations obtained from the search.

Evaluation Metrics. We evaluate the performance of each method on two folds, model performance and computational efficiency. The performance depends on the downstream applications with common metrics such as accuracy for image classification and EM or F1-scores for question and answering tasks. Computational efficiency is assessed by BOPs, where lower values indicate more compact models with typically faster inference. For ease of comparison, we report the relative BOP ratio against the baseline full precision models.

6.1. CNN Architectures

ResNet20 on CIFAR10. We first test our framework **GETA** using ResNet20 on CIFAR10 dataset. For fair comparison, only weight quantization is applied, excluding activation quantization. As shown in Tab. 2, **GETA** achieves a 4.5% relative BOPs compression ratio with only a loss of 0.28% in test accuracy, which demonstrates significantly better performance than ANNC [61]. Compared to QST-B [47], **GETA** reduces BOP by 13% with only a minimal accuracy drop of 0.08%. We argue that **GETA** is better suited for practical applications, as QST-B focuses on joint unstructured pruning and quantization. While unstructured pruning tends to deliver higher accuracy at similar compression ratios, its theoretical speedups are challenging to achieve without specialized hardware and software supports [24, 30, 66]. In contrast, the structurally pruned and quantized model produced by **GETA** is more easily deployed in practical applications.

VGG7 on CIFAR10. We then test **GETA** using VGG7 on CIFAR10 to compare with the joint structured pruning and quantization benchmarks. In this case, we enable both weight and activation quantization and report the results in Tab. 3. Based on the results, **GETA** could significantly outperform other competitors in terms of the test accuracy

Table 2. ResNet20 on CIFAR10 dataset. The red and orange represent the best and second-best results, respectively, in the last two columns. Same rule is followed in Tab. 3 and Tab. 4.

Method	Pruning	Wt Quant	Act Quant	Accuracy (%)	Rel. BOPs (%)
Baseline	✗	✗	✗	91.70	100
ANNC [61]	Unstructured	✓	✗	90.90	6.1
QST-B [47]	Unstructured	✓	✗	91.50	5.1
GETA	Structured	✓	✗	91.42	4.5

by 0.61% - 1.14%, and achieves the second best relative BOP ratio which is only worse than BB [55]. BB separates the model architecture compression and training stages, requiring substantial effort for each. In contrast, **GETA** offers practical advantages, including efficiency and broad architecture compatibility, enabling an end-to-end, automated joint structured pruning and quantization approach.

Table 3. VGG7 on CIFAR10 dataset.

Method	Pruning	Wt Quant	Act Quant	Accuracy (%)	Rel. BOPs (%)
Baseline	✗	✗	✗	93.05	100
DJPQ [58]	Structured	✓	✓	91.54	0.48
DJPQ-restrict [58]	Structured	✓	✓	91.43	0.46
BB [55]	Structured	✓	✓	91.96	0.29
GETA	Structured	✓	✓	92.57	0.41

ResNet50 on ImageNet. We next test **GETA** using ResNet50 on ImageNet. We select ResNet50 on ImageNet because it serves as one of most common benchmarks in structured pruning works [18, 39], while studies on joint structured pruning and quantization seem absent to the best of our knowledge. Therefore, we compare with joint unstructured pruning or semi-structured pruning and quantization works OBC [19] and Clip-Q [52]. Unlike the CIFAR10 experiments, we start the training from a pretrained checkpoint. As the results present in Tab. 4, **GETA** could consistently outperform them in terms of both test accuracy and relative BOP ratios. Considering the difficulty of performance preservation for structured pruning, **GETA** demonstrates superior performance to existing works.

Table 4. ResNet50 on ImageNet dataset.

Method	Pruning	Wt Quant	Act Quant	Accuracy (%)	Rel. BOPs (%)
Baseline	✗	✗	✗	76.13	100
OBC [19]	Semi-Structured	✓	✗	71.47	6.67
Clip-Q [52]	Unstructured	✓	✗	73.70	6.30
GETA (40% sparsity)	Structured	✓	✗	75.10	6.97
GETA (50% sparsity)	Structured	✓	✗	74.40	5.38

6.2. Transformer

Bert on SQuAD. We now apply **GETA** to the transformer architecture. The first is the representative encoder-based BERT model [56] on the SQuAD benchmark [49]. While previous works on joint quantization and structured pruning have not been applied to the transformer architecture, we

Table 5. Comparison of **GETA** vs. Structured Pruning followed by Post-Training Quantization (PTQ) for BERT on SQuAD.

Method	Sparsity	EM (%)	F1 (%)	BOPs (GB)	Rel. BOPs (%)
Baseline	0%	80.08	88.50	13.57	100.0
OTO [10] followed up 8-bit PTQ	10%	73.87	83.43	3.17	23.4
	30%	72.95	83.31	2.71	20.0
	50%	72.71	83.30	2.26	16.7
	70%	71.24	82.57	1.80	13.3
	10%	78.26	86.06	2.63	19.4
GETA	30%	77.28	85.70	2.29	16.9
	50%	76.74	85.87	1.96	14.4
	70%	75.88	84.74	1.62	11.9

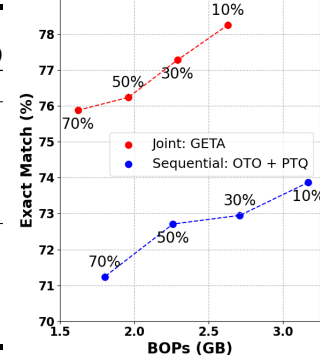


Figure 3. Phi2-2.7B.

make a more relevant comparison by contrasting our joint optimization approach with the sequential baseline, which first applies pruning-aware training (HESSO) [12] and then performs standard post-training quantization (PTQ) [48]. An alternative sequential baseline, the quantize-then-prune approach, is excluded from our comparison for the following two reasons: (i) Applying PTQ to the full model introduces challenges when attempting to prune the model afterward, as calculating gradients with respect to quantized values requires careful handling. (ii) A recent work [25] mathematically shows that prune-then-quantize approach is the optimal sequential strategy. Therefore, we focus on comparing **GETA** with the prune-then-quantize baselines.

The comparison in Tab. 5 clearly highlights the advantages of joint structured pruning and quantization during training, versus only pruning at training time and quantization during post-training. At all sparsity ratios, **GETA** consistently outperforms the multi-stage approach by a large margin. In particular, we observe improvements in exact-match rates (EM) and F1-scores while achieving better compression rates. These results empirically validate that joint pruning and quantization during training is superior to the conventional approach of pruning-aware training followed by post-training quantization, both in terms of model quality and computational efficiency.

Phi2 on Common-Sense. We next evaluate **GETA** on popular large language models. Since **GETA** leverages full gradient information, we select Phi2-2.7B [32], a model with fewer than 3 billion parameters, to ensure computational feasibility on a single A100 GPU. Similar to the experiments on BERT, we compare **GETA** with a prune-then-quantize baseline. This baseline first applies pruning-aware training techniques, including SliceGPT [2], LoraShear [8], LoraPrune [65], and LLMPruner [43], followed by PTQ. For a fair comparison, the average bit width

across all layers after applying **GETA** is set to approximately 8 bits, while the baseline uses uniform 8-bit PTQ. As shown in Fig. 3, **GETA** consistently outperforms all prune-then-quantize baselines in terms of average performance in common-sense tasks including BoolQ, PIQA, HellaSwag, WinoGrande, ARC-e, ARC-c and OBQA.

Vision Transformers. Finally, we evaluate **GETA** on a variety of vision transformer architectures, including SimpleViT [60], ViT [1], DeiT [51], Swin Transformer [41], and Pyramid Vision Transformer [57]. These models are selected to further validate the architecture-agnostic nature of the **GETA** framework. To highlight this capability, we focus on reporting the test accuracy and relative BOPs compared to the baseline models. The promising results, as shown in Tab. 6, demonstrate the efficiency and versatility of **GETA** across diverse transformer architectures.

Table 6. Experiments on various vision transformer architectures.

Dataset	Model	Base Acc (%)	Acc (%)	Rel. BOPs (%)
Cifar10	SimpleViT	86.48	86.06	4.95
ImageNet	ViT-Small	81.43	80.12	19.37
	DeiT-Tiny	72.01	72.88	16.95
	Swin-Tiny	80.92	80.09	21.84
	PVTv2-B2	81.69	80.53	17.39

7. Conclusion

We proposed **GETA**, an automatic framework designed to jointly apply structured pruning and quantization-aware training to deep neural networks, addressing key limitations of existing methods. By leveraging quantization-aware dependency graph analysis, **GETA** supports a wide range of architectures. The proposed QASSO optimizer provides explicit control over bit width and sparsity, resolving black-box issues in existing approaches. With improved generalization, white-box optimization, and a one-shot framework, **GETA** offers an easy-to-use and user-friendly solution for practical deployment. In the future, it will be interesting to explore adapting **GETA** for specialized hardware to improve real-world deployment on different platforms.

Acknowledgments. We acknowledge the Microsoft Applied Sciences Group for their computational support.

References

- [1] Dosovitskiy Alexey. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 7, 8
- [2] Saleh Ashkboos, Maximilian L Croci, Marcelo Gennari do Nascimento, Torsten Hoefer, and James Hensman. Slicept: Compress large language models by deleting rows and columns. In *International Conference on Learning Representations*, 2024. 8
- [3] Konstantinos Balaskas, Andreas Karatzas, Christos Sad, Kostas Siozios, Iraklis Anagnostopoulos, Georgios Zervakis, et al. Hardware-aware DNN compression via diverse pruning and mixed-precision quantization. *IEEE Transactions on Emerging Topics in Computing*, 2024. 1, 2, 3
- [4] Dimitri P Bertsekas. Nonlinear programming. *Journal of the Operational Research Society*, 48(3):334–334, 1997. 5
- [5] Tianyi Chen, Frank E Curtis, and Daniel P Robinson. A reduced-space algorithm for minimizing ℓ_1 -regularized convex functions. *SIAM Journal on Optimization*, 27(3):1583–1610, 2017. 3
- [6] Tianyi Chen, Frank E Curtis, and Daniel P Robinson. Farsa for ℓ_1 -regularized convex optimization: local convergence and numerical experience. *Optimization Methods and Software*, 2018. 3
- [7] Tianyi Chen, Tianyu Ding, Bo Ji, Guanyi Wang, Yixin Shi, Jing Tian, Sheng Yi, Xiao Tu, and Zhihui Zhu. Orthant based proximal stochastic gradient method for ℓ_1 -regularized optimization. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2020, Ghent, Belgium, September 14–18, 2020, Proceedings, Part III*, pages 57–73. Springer, 2021. 3
- [8] Tianyi Chen, Tianyu Ding, Badal Yadav, Ilya Zharkov, and Luming Liang. Lorashear: Efficient large language model structured pruning and knowledge recovery. *arXiv preprint arXiv:2310.18356*, 2023. 8
- [9] Tianyi Chen, Tianyu Ding, Zhihui Zhu, Zeyu Chen, Hsiang-Tao Wu, Ilya Zharkov, and Luming Liang. Otov3: Automatic architecture-agnostic neural network training and compression from structured pruning to erasing operators. *arXiv preprint arXiv:2312.09411*, 2023. 3
- [10] Tianyi Chen, Bo Ji, Tianyu Ding, Biyi Fang, Guanyi Wang, Zhihui Zhu, Luming Liang, Yixin Shi, Sheng Yi, and Xiao Tu. Only train once (oto): A one-shot neural network training and pruning framework. *Neurips 2021*, 2021. 1, 3, 8
- [11] Tianyi Chen, Luming Liang, Ilya Zharkov Tianyu Ding, and Zhihui Zhu. Otov2: Automatic, generic, user-friendly. *ICLR 2023*, 2023. 1, 3, 4
- [12] Tianyi Chen, Xiaoyi Qu, David Aponte, Colby Banbury, Jongwoo Ko, Tianyu Ding, Yong Ma, Vladimir Lyapunov, Ilya Zharkov, and Luming Liang. Hesso: Towards automatic efficient and user friendly any neural network training and pruning. *arXiv preprint arXiv:2409.09085*, 2024. 3, 5, 6, 8
- [13] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 7
- [14] Lei Deng, Guoqi Li, Song Han, Luping Shi, and Yuan Xie. Model compression and hardware acceleration for neural networks: A comprehensive survey. *Proceedings of the IEEE*, 108(4):485–532, 2020. 1
- [15] Zhen Dong, Zhewei Yao, Daiyaan Arfeen, Amir Gholami, Michael W Mahoney, and Kurt Keutzer. Hawq-v2: Hessian aware trace-weighted quantization of neural networks. *Advances in neural information processing systems*, 33:18518–18529, 2020. 3
- [16] Zhen Dong, Zhewei Yao, Amir Gholami, Michael W Mahoney, and Kurt Keutzer. Hawq: Hessian aware quantization of neural networks with mixed-precision. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 293–302, 2019. 3
- [17] Ahmed T Elthakeb, Prannoy Pilligundla, Fatemehsadat Miresghallah, Amir Yazdanbakhsh, and Hadi Esmaeilzadeh. Releq: A reinforcement learning approach for automatic deep quantization of neural networks. *IEEE micro*, 40(5):37–45, 2020. 3
- [18] Gongfan Fang, Xinyin Ma, Mingli Song, Michael Bi Mi, and Xinchao Wang. Depgraph: Towards any structural pruning. *arXiv preprint arXiv:2301.12900*, 2023. 1, 3, 7
- [19] Elias Frantar and Dan Alistarh. Optimal brain compression: A framework for accurate post-training quantization and pruning. *Advances in Neural Information Processing Systems*, 35:4475–4488, 2022. 7
- [20] Elias Frantar and Dan Alistarh. Sparsegpt: Massive language models can be accurately pruned in one-shot. In *International Conference on Machine Learning*, pages 10323–10337. PMLR, 2023. 3
- [21] Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac’h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. A framework for few-shot language model evaluation, 07 2024. 7
- [22] Shang-Hua Gao, Yong-Qiang Tan, Ming-Ming Cheng, Chengze Lu, Yunpeng Chen, and Shuicheng Yan. Highly efficient salient object detection with 100k parameters. In *European Conference on Computer Vision*, pages 702–721. Springer, 2020. 3
- [23] Song Han, Huizi Mao, and William J Dally. Deep compression: Compressing deep neural networks with pruning, trained quantization and Huffman coding. *arXiv preprint arXiv:1510.00149*, 2015. 1
- [24] Cong Hao, Xiaofan Zhang, Yuhong Li, Sitao Huang, Jinjun Xiong, Kyle Rupnow, Wen-mei Hwu, and Deming Chen. FPGA/DNN co-design: An efficient design methodology for IoT intelligence on the edge. In *Proceedings of the 56th Annual Design Automation Conference 2019*, pages 1–6, 2019. 7
- [25] Simla Burcu Harma, Ayan Chakraborty, Elizaveta Kostenok, Danila Mishin, Dongho Ha, Babak Falsafi, Martin Jaggi, Ming Liu, Yunho Oh, Suvinay Subramanian, et al. Effective interplay between sparsity and quantization: From theory to practice. *ICLR 2025*, 2025. 8

- [26] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016. 1, 7
- [27] Yang He and Lingao Xiao. Structured pruning for deep convolutional neural networks: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 2023. 14
- [28] Cheeun Hong, Sungyong Baik, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Cadyq: Content-aware dynamic quantization for image super-resolution. In *European Conference on Computer Vision*, pages 367–383. Springer, 2022. 3
- [29] Peng Hu, Xi Peng, Hongyuan Zhu, Mohamed M Sabry Aly, and Jie Lin. Opq: Compressing deep neural networks with one-shot pruning-quantization. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 7780–7788, 2021. 1, 3
- [30] Sitao Huang, Carl Pearson, Rakesh Nagi, Jinjun Xiong, Deming Chen, and Wen-mei Hwu. Accelerating sparse deep neural networks on fpgas. In *2019 IEEE High Performance Extreme Computing Conference (HPEC)*, pages 1–7, 2019. 7
- [31] Yerlan Idelbayev and Miguel Á Carreira-Perpiñán. Exploring the effect of ℓ_0/ℓ_2 regularization in neural network pruning using the lc toolkit. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3373–3377. IEEE, 2022. 3
- [32] Mojan Javaheripi, Sébastien Bubeck, Marah Abidin, Jyoti Aneja, Sebastian Bubeck, Caio César Teodoro Mendes, Weizhu Chen, Allie Del Giorno, Ronen Eldan, Sivakanth Gopi, et al. Phi-2: The surprising power of small language models. *Microsoft Research Blog*, 1(3):3, 2023. 7, 8
- [33] Simonyan Karen. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 7
- [34] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186, 2019. 1, 7
- [35] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. 7
- [36] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012. 1
- [37] Yunsong Li, Xin Zhang, Weiying Xie, Jiaqing Zhang, Leyuan Fang, and Jiawei Du. Markov-pq: Joint pruning-quantization via learnable markov chain. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024. 1, 2
- [38] Shaohui Lin, Rongrong Ji, Yuchao Li, Cheng Deng, and Xuelong Li. Toward compact convnets via structure-sparsity regularized filter pruning. *IEEE transactions on neural networks and learning systems*, 31(2):574–588, 2019. 3
- [39] Shaohui Lin, Rongrong Ji, Chenqian Yan, Baoshang Zhang, Liujuan Cao, Qixiang Ye, Feiyue Huang, and David Doermann. Towards optimal structured cnn pruning via generative adversarial learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2790–2799, 2019. 7
- [40] Jing Liu, Bohan Zhuang, Peng Chen, Chunhua Shen, Jianfei Cai, and Minghui Tan. Single-path bit sharing for automatic loss-aware model compression. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(10):12459–12473, 2023. 3
- [41] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 8
- [42] Christos Louizos, Karen Ullrich, and Max Welling. Bayesian compression for deep learning. In *Advances in neural information processing systems*, pages 3288–3298, 2017. 1
- [43] Xinyin Ma, Gongfan Fang, and Xinchao Wang. Llm-pruner: On the structural pruning of large language models. *Advances in neural information processing systems*, 36:21702–21720, 2023. 8
- [44] Fanxu Meng, Hao Cheng, Ke Li, Huixiang Luo, Xiaowei Guo, Guangming Lu, and Xing Sun. Pruning filter in filter. *arXiv preprint arXiv:2009.14410*, 2020. 3
- [45] Srinivas S Miriyala, PK Suhas, Utsav Tiwari, and Vikram N Rajendiran. Mixed precision neural quantization with multi-objective bayesian optimization for on-device deployment. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6260–6264. IEEE, 2024. 3
- [46] Jorge Nocedal and Stephen J Wright. *Numerical optimization*. Springer, 1999. 5
- [47] Jun-Hyung Park, Kang-Min Kim, and Sangkeun Lee. Quantized sparse training: A unified trainable framework for joint pruning and quantization in dnns. *ACM Transactions on Embedded Computing Systems*, Vol. 21, No. 5, Article 60, 2022. 1, 2, 3, 7
- [48] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, 2019. 8
- [49] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*, 2016. 7
- [50] Md Maruf Hossain Shuvo, Syed Kamrul Islam, Jianlin Cheng, and Bashir I Morshed. Efficient acceleration of deep learning inference on resource-constrained edge devices: A review. *Proceedings of the IEEE*, 111(1):42–91, 2022. 1
- [51] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pages 10347–10357. PMLR, 2021. 8

- [52] Frederick Tung and Greg Mori. Clip-q: Deep network compression learning by in-parallel pruning-quantization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7873–7882, 2018. 1, 2, 3, 7
- [53] Stefan Uhlich, Lukas Mauch, Fabien Cardinaux, Kazuki Yoshiyama, Javier Alonso Garcia, Stephen Tiedemann, Thomas Kemp, and Akira Nakamura. Mixed precision dnns: All you need is a good parametrization. In *International Conference on Learning Representations*, 2020. 2, 3
- [54] Stefan Uhlich, Lukas Mauch, Kazuki Yoshiyama, Fabien Cardinaux, Javier Alonso Garcia, Stephen Tiedemann, Thomas Kemp, and Akira Nakamura. Differentiable quantization of deep neural networks. *arXiv preprint arXiv:1905.11452*, 2(8), 2019. 3
- [55] Mart van Baalen, Christos Louizos, Markus Nagel, Rana Ali Amjad, Ying Wang, Tijmen Blankevoort, and Max Welling. Bayesian bits: Unifying quantization and pruning. *arXiv preprint arXiv:2005.07093*, 2020. 1, 2, 3, 7
- [56] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 1, 7
- [57] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pvt v2: Improved baselines with pyramid vision transformer. *Computational Visual Media*, 8(3):415–424, 2022. 8
- [58] Ying Wang, Yadong Lu, and Tijmen Blankevoort. Differentiable joint pruning and quantization for hardware efficiency. In *European Conference on Computer Vision*, pages 259–277. Springer, 2020. 1, 2, 3, 7
- [59] Wei Wen, Chunpeng Wu, Yandan Wang, Yiran Chen, and Hai Li. Learning structured sparsity in deep neural networks. *arXiv preprint arXiv:1608.03665*, 2016. 3
- [60] Weiyang Xie, Haowei Li, Jitao Ma, Yunsong Li, Jie Lei, Donglai Liu, and Leyuan Fang. Jointsq: Joint sparsification-quantization for distributed learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5778–5787, 2024. 8
- [61] Haichuan Yang, Shupeng Gui, Yuhao Zhu, and Ji Liu. Automatic neural network compression by sparsity-quantization joint learning: A constrained optimization-based approach. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2178–2188, 2020. 1, 2, 3, 7
- [62] Huanrui Yang, Wei Wen, and Hai Li. Deepfoyer: Learning sparser neural network with differentiable scale-invariant sparsity measures. In *International Conference on Learning Representations*, 2020. 3
- [63] Zhewei Yao, Zhen Dong, Zhangcheng Zheng, Amir Gholami, Jiali Yu, Eric Tan, Leyuan Wang, Qijing Huang, Yida Wang, Michael Mahoney, et al. Hawq-v3: Dyadic neural network quantization. In *International Conference on Machine Learning*, pages 11875–11886. PMLR, 2021. 3
- [64] Shaokai Ye, Tianyun Zhang, Kaiqi Zhang, Jiayu Li, Jiaming Xie, Yun Liang, Sijia Liu, Xue Lin, and Yanzhi Wang. A unified framework of dnn weight pruning and weight clustering/quantization using admm. *arXiv preprint arXiv:1811.01907*, 2018. 1, 3
- [65] Mingyang Zhang, Hao Chen, Chunhua Shen, Zhen Yang, Linlin Ou, Xinyi Yu, and Bohan Zhuang. Loraprune: Pruning meets low-rank parameter-efficient fine-tuning. *arXiv preprint arXiv:2305.18403*, 2023. 8
- [66] Xiaofan Zhang, Xinheng Liu, Anand Ramachandran, Chuanhao Zhuge, Shibin Tang, Peng Ouyang, Zuofu Cheng, Kyle Rupnow, and Deming Chen. High-performance video content recognition with long-term recurrent convolutional network for fpga. In *2017 27th International Conference on Field Programmable Logic and Applications (FPL)*, pages 1–4. IEEE, 2017. 7
- [67] Yuefu Zhou, Ya Zhang, Yanfeng Wang, and Qi Tian. Accelerate cnn via recursive bayesian pruning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3306–3315, 2019. 3
- [68] Tao Zhuang, Zhixuan Zhang, Yuheng Huang, Xiaoyi Zeng, Kai Shuang, and Xiang Li. Neuron-level structured pruning using polarization regularizer. *Advances in Neural Information Processing Systems*, 33, 2020. 3