

Layer-Wise High-Impact Parameter Ratio Optimization in Post-Training Quantization for Large Language Models

Cuong Pham¹ Hoang Anh Dung¹ Cuong C. Nguyen²
Trung Le¹ Gustavo Carneiro² Thanh-Toan Do¹

¹Department of Data Science and AI, Monash University, Australia

²Centre for Vision, Speech and Signal Processing, University of Surrey, UK

Abstract

Large language models (LLMs) have advanced natural language processing, but their massive parameter counts create computational and memory challenges during deployment. Post-training quantization (PTQ) has emerged as a promising approach to mitigate these challenges. While existing PTQ methods can effectively quantize LLMs, they experience substantial accuracy loss at extremely low bit-widths due to high-impact parameters. Several approaches address this by retaining high-impact parameters in FP16 format, but they apply fixed ratios across all layers, overlooking layer-wise sensitivity variations. We propose a quadratic optimization framework that determines layer-specific ratios of high-impact parameters while considering inter-layer dependencies. We quantize high-impact parameters to moderate bit-widths while the remaining parameters are quantized to extremely low bit-widths. Under the same resource budget, this preserves more high-impact parameters than methods retaining a few in FP16 format. Our framework enables leveraging advanced quantization methods for high-impact parameters while applying lightweight computational quantization methods to the rest, achieving an effective balance between computational efficiency and accuracy during quantization process.

1 Introduction

Large language models (LLMs) (Wei et al., 2022a; Touvron et al., 2023; Zhang et al., 2022), have gained significant attention due to their remarkable performance in handling complex natural language tasks (Hendrycks et al., 2021), such as language generation, translation, question answering, and text summarization. However, the significantly large size of these models, often comprising billion-level parameters, demands significant computational resources for inference and deployment. To address this issue, an attractive approach is network quantization (Frantar et al., 2023), which not

only reduces the computation costs, but also significantly reduces the memory usage. Compared to quantization aware training (QAT) that requires huge training costs and access to large training data, the post-training quantization (PTQ) (Frantar et al., 2023; Lin et al., 2024; Xiao et al., 2023; Wei et al., 2023), which requires limited calibration data and computational resources, is more in demand for quantizing LLMs.

Previous PTQ methods (Wei et al., 2023; Xiao et al., 2023; Shao et al., 2024) for LLMs have demonstrated the ability to quantize LLMs to lower precision. For example, quantizing weights to 4 bits often results in negligible performance degradation. However, when quantized to extremely low bit-width (e.g., 2-bit weight-only quantization), these methods lead to significant performance degradation compared to full-precision models. Several approaches (Lin et al., 2024; Kim et al., 2024; Shao et al., 2024; Cui and Wang, 2024) indicate that high-impact parameters that significantly influence quantization performance are often the biggest challenge in LLM quantization. In AWQ (Lin et al., 2024), the authors indicate that not all weights in an LLM are equally important, and protecting only a small percentage of high-impact weights can greatly reduce quantization error. In CherryQ (Cui and Wang, 2024), the authors separate high-impact parameters in an element-wise manner. In SqueezeLLM (Kim et al., 2024), the authors represent sensitive weight values in an efficient sparse format and adopt codebook-based non-uniform quantization, which is not hardware-friendly compared to standard uniform quantization methods. Although these approaches successfully improve performance of PTQ for LLMs, applying a fixed ratio of high-impact parameters across all layers overlooks layer-wise variation in parameter importance. For instance, in LLaMA-2-7B, the self-attention key projection at block 2 contains a large fraction of channels with high Fisher information,

whereas the MLP down-projection at block 3 contains only a small number of high-impact channels (see Figure 1). Consequently, this approach may be suboptimal as the ratio of importance of parameters differs across layers. Additionally, handling important parameters in an element-wise manner in CherryQ (Cui and Wang, 2024) or using non-uniform quantization as in SqueezeLLM (Kim et al., 2024) may not be hardware-friendly and pose challenges for hardware deployment.

In this paper, we propose a principled approach inspired by (Chen et al., 2021; Deng et al., 2023) for mixed-precision quantization of convolutional neural networks via quadratic optimization to identify the accurate ratio of high-impact parameters for each layer in a channel-wise manner. Specifically, we propose an efficient quadratic optimization approach to determine the optimal ratio of high-impact parameters, which takes into account the impact of different layer sensitivities on model performance. Under the same resource-constrained budget, retaining all high-impact parameters in FP16 format or full precision is not always the best strategy, as it limits the total number of such parameters that can be preserved due to high precision cost. Instead, quantizing high-impact parameters to moderate bit-widths (e.g., 4-bit), the precision that results in negligible performance degradation in quantized models, allows preserving a greater ratio of high-impact parameters for each layer within the budget. Meanwhile, the remaining normal parameters can be quantized to extremely low bit-widths (e.g., 2-bit). Additionally, the more advanced and accurate quantization methods such as AdaRound (Nagel et al., 2020) often lead to better overall results. However, the number of learnable parameters when applying AdaRound is equal to the number of weight parameters, which makes it infeasible to apply AdaRound directly to large language models. To address this, CBQ (Ding et al., 2025) proposes LoRA-Rounding, which considerably reduces the number of learnable parameters. Although this technique achieves substantial improvements for quantizing LLMs, it depends on rank selection, and the significantly reduced parameter space may constrain the search for optimal rounding solutions. Therefore, different from previous methods (Cui and Wang, 2024; Kim et al., 2024; Ding et al., 2025), the proposed approach allows us to adopt a hybrid quantization strategy to balance accuracy and computational efficiency by leveraging advanced quantization methods such

as AdaRound only for high-impact parameters, while quantizing the vast majority of remaining parameters using more lightweight quantization approaches such as OmniQuant (Shao et al., 2024). This hybrid strategy effectively allocates computational resources based on optimized high-impact parameter allocation, significantly reducing overhead and maintaining overall model quality.

To summarize, the main contributions of this paper are as follows: We propose a novel approach to determine the optimal ratios of high-impact parameters across layers in large language models. Rather than applying a fixed ratio, our approach captures layer-wise sensitivity variations through a quadratic optimization framework. This enables more accurate allocation of quantization precision based on each layer’s contribution to overall model performance. Additionally, the proposed approach allows us to utilize advanced PTQ methods that require optimizing a large learnable matrix, such as AdaRound (Nagel et al., 2020), for only the most important parameters. By combining an advanced PTQ method that optimizes a learnable weight-rounding matrix for the most impactful parameters with a lightweight quantization method for the remaining parameters, our approach achieves a strong balance between performance and computational efficiency during the quantization process through optimized high-impact parameter allocation.

2 Related works

Quantization for large language models. Quantization is an effective approach to compress LLMs, which helps to significantly reduce inference cost and memory usage. Many studies (Liu et al., 2023b; Shao et al., 2024; Wei et al., 2022c, 2023) show the presence of high-impact parameters in LLMs, and these parameters make the quantization process more challenging and require special handling. One common approach is to use an equivalent transformation to handle the high-impact parameters in LLMs. Other approaches (Xiao et al., 2023; Shao et al., 2024; Lin et al., 2024) propose shifting transformations and scaling transformations to further improve the performance of quantized models. Recently, rotation transformations (Tseng et al., 2024; Ashkboos et al., 2024; Liu et al., 2025) have been proposed to handle high-impact parameters in LLMs. Several works also explore mixed-precision quantization (Kim et al., 2024; Lee et al., 2024; Cui and Wang, 2024) to achieve

a better trade-off between accuracy and efficiency. Learning-based mixed-precision methods such as HAQ (Wang et al., 2019) and HAWQ (Dong et al., 2019) search for per-layer bit-widths using reinforcement learning or second-order information. In contrast, our method operates at channel granularity within a fixed bit-width, optimizing the per-layer ratio of high-impact parameters rather than the bit-width itself.

Handling high-impact parameters for PTQ in LLMs using mixed-precision quantization. As previously mentioned, high-impact parameters are widely found in the activations and weights of large language models, posing a significant challenge for quantization. Consequently, many mixed-precision methods for LLMs aim to represent a small number of high-impact parameters in higher precision while quantizing other values in lower precision. LLM.int8() (Dettmers et al., 2022) introduces mixed-precision decomposition to handle high-impact features, which are specific hidden dimensions with extremely large magnitudes that appear systematically across transformer layers. In CherryQ (Cui and Wang, 2024), the authors highlight the importance of cherry parameters, which have a significant impact on quantization performance. They measure parameter impact and identify the most important of parameters for each layer in element-wise manner. They then keep these important parameters in FP16 and then fine-tune the model using quantization-aware training. In SqueezeLLM (Kim et al., 2024), the authors use non-uniform quantization to assign higher bit-widths to more important parameters. While they achieve substantial performance improvements in extreme low-bitwidth quantization, applying a uniform ratio of high-impact parameters across all layers does not take into account the variation in parameter importance throughout the network, potentially leading to suboptimal allocation of precision resources. In contrast, our method addresses high-impact parameters by taking into account the layer-wise dependencies of the models.

Advanced quantization for quantizing large language models. In standard PTQ, rounding-to-nearest is the most common approach as it minimizes quantization error in weight space. However, recent state-of-the-art methods (Nagel et al., 2020; Li et al., 2021; Liu et al., 2023a; Wei et al., 2022b; Jeon et al., 2023; Pham et al., 2024) demonstrate that optimizing the rounding function with respect to the task loss significantly improves quan-

tization performance. AdaRound (Nagel et al., 2020) introduces a differentiable rounding mechanism that adapts to data distribution and task objectives, showing remarkable performance improvements for convolutional neural networks (Wei et al., 2022b; Lin et al., 2023; Jeon et al., 2023), diffusion models (Shang et al., 2023; Li et al., 2023; Huang et al., 2024), and LLMs (Ding et al., 2025). However, applying AdaRound directly to LLMs presents significant challenges, as the number of learnable parameters equals the number of weights in the model, making optimization computationally prohibitive at scale. To overcome this challenge, CBQ (Ding et al., 2025) introduces LoRA-Rounding, which considerably reduces the quantity of parameters that need to be learned. Although this technique achieves substantial improvements for quantizing LLMs, it depends on rank selection, and the significantly reduced parameter space may constrain the search for optimal rounding solutions. In this work, the proposed approach to determine the optimal ratio of high-impact parameters for each layer allows us to take advantage of AdaRound’s quantization capabilities while maintaining computational efficiency when quantizing LLMs.

3 Proposed method

3.1 Preliminary analysis

To examine the sensitivity distribution across hidden dimensions (channels) in large language models (LLMs), we compute the average Fisher information of the parameters associated with each hidden dimension and use it as an importance metric. This metric captures how much each dimension contributes to model performance under quantization. Our experiments employ the LLaMA-2-7B model with 128 calibration samples from WikiText-2. As shown in Figure 1, the importance scores are significantly different across hidden dimensions, showing that influential parameters are concentrated in a few specific dimensions. This observation shows that we can adopt channel-wise strategy for mitigating problems in handling high-impact parameters, which offers greater hardware efficiency than the element-wise methods used in prior work (Cui and Wang, 2024). Moreover, we observe substantial layer-to-layer variability in the distribution of high-impact parameters—certain layers contain more influential parameters than others, which indicates that applying fixed ratios of high-impact parameters across all layers may not be optimal.

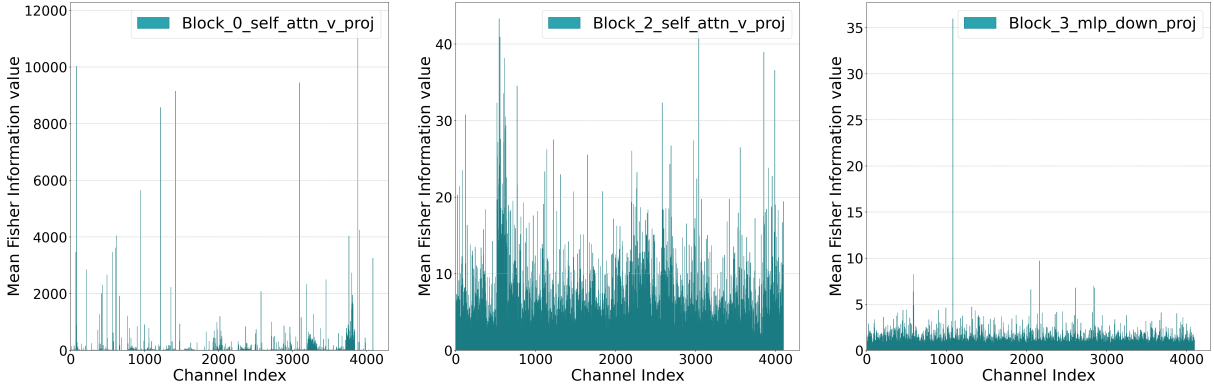


Figure 1: Distribution of average Fisher information across channel dimensions at several layers in the LLaMA-2-7B model. There is a significant difference in importance scores for each layer, indicating that impactful parameters are concentrated in a subset of channels. Additionally, the ratio of high-impact parameters differs across layers.

3.2 Determining optimal ratio of high-impact parameters

Let θ_{FP} denote the model’s full-precision weights and θ_Q denote the quantized weights. Our goal is to identify the most high-impact parameters in the model to quantize at higher precision, while quantizing the remaining parameters at lower precision. The quantization process for a network’s weights can be expressed as $\theta_Q = \text{Quant}(\theta_{FP}) = \theta_{FP} + \Delta$, where $\text{Quant}(\cdot)$ denotes the quantization function and Δ is the quantization error.

Given the full-precision weights θ_{FP} , the training set $X^{(T)} = \{x_i\}_{i=1}^N$ and the model weight quantization error $\Delta \in \mathbb{R}^{|\theta_{FP}| \times 1}$, the change in the loss of the quantized model can be approximated using a second-order Taylor expansion around θ_{FP} as follows:

$$\begin{aligned} \mathcal{L}(\theta_{FP} + \Delta, X^{(T)}) - \mathcal{L}(\theta_{FP}, X^{(T)}) \\ = g^\top \Delta + 1/2 \Delta^\top \mathbf{H} \Delta + \mathcal{O}(\Delta^\top \Delta), \end{aligned} \quad (1)$$

where $g = \frac{1}{N} \sum_{i=1}^N \nabla_{\theta_{FP}} [\mathcal{L}(\theta_{FP}, x_i)]$ denotes the expected gradient of the loss with respect to θ_{FP} , and $\mathbf{H} = \frac{1}{N} \sum_{i=1}^N \nabla_{\theta_{FP}}^2 [\mathcal{L}(\theta_{FP}, x_i)]$ denotes the Hessian matrix. For a well-trained model, $\|g\| \approx 0$, we can simplify the Eq. (1) to:

$$\begin{aligned} \mathcal{L}(\theta_{FP} + \Delta, X^{(T)}) - \mathcal{L}(\theta_{FP}, X^{(T)}) \\ \approx 1/2 \Delta^\top \mathbf{H} \Delta. \end{aligned} \quad (2)$$

In the context of quantization, we define $\mathcal{L}(\theta_Q, X^{(T)}) = \mathcal{L}(\theta_{FP} + \Delta, X^{(T)})$ as the reconstruction loss between the output of the quantized model and that of the full-precision model as follows:

$$\mathcal{L}(\theta_Q, X^{(T)}) = \frac{1}{N} \sum_{i=1}^N \|f_{\theta_{FP}}(x_i) - f_{\theta_Q}(x_i)\|_F^2, \quad (3)$$

where f is the output of the model of interest.

Measure the impact of weight parameters on model performance. In post-training quantization, a principled approach to assessing parameters’ importance is to measure the sensitivity of the training loss to perturbations in individual weight values. This sensitivity is formally captured by the diagonal entries of the Hessian matrix of the loss function with respect to the model parameters. For a given parameter $\theta_{FP,i}$, the corresponding entry $\mathbf{H}_{i,i} = \mathbb{E}_{x_j \sim X^{(T)}} \left[\nabla_{\theta_{FP}}^2 [\mathcal{L}(\theta_{FP}, x_i)] \right]_{i,i}$ represents the expected second-order derivative of the loss, and thus characterizes the local curvature of the loss landscape around the full-precision solution. A larger value of $\mathbf{H}_{i,i}$ implies that the loss is more sensitive to small deviations in $\theta_{FP,i}$, indicating a higher risk of performance degradation under quantization. Therefore, parameters with higher curvature should be preserved with greater precision.

We refer to $\mathbf{H}_{i,i}$ as the impact score of the parameter $\theta_{FP,i}$, as it reflects the parameter’s relative importance in maintaining model fidelity after quantization.

We approximate the Hessian matrix \mathbf{H} using the Fisher Information Matrix (\mathbf{F}) as follows:

$$\mathbf{H}_{i,i} \approx \mathbf{F}_{i,i} = \mathbb{E} \left[gg^\top \right]_{i,i}. \quad (4)$$

Optimal ratio for high-impact parameters. Following the estimation of the parameter impact,

we rank all weight parameters according to their scores. Based on this ranking, we designate a subset of parameters with the highest impact as high-impact parameters, which are more susceptible to quantization error. To preserve model performance, these high-impact parameters are quantized using a higher precision bit-width, denoted as b^H , while the remaining parameters are quantized using a standard lower bit-width b^N . The proportion of parameters assigned to b^H is treated as an optimization objective and is selected based on the optimization of Eq. (2). Given that the model has L layers and there are $|\mathbb{B}|$ possible options for the high-impact parameter ratio in each layer (where \mathbb{B} is the set of candidate ratios), we have $|\mathbb{B}|^L$ potential configurations of the high-impact parameter ratios across all layers. Because exhaustive search over these $|\mathbb{B}|^L$ configurations is infeasible, we cast the selection as a quadratic optimization problem, derived below.

For each layer l , we define a one-hot vector $\delta_l \in \{0, 1\}^{|\mathbb{B}|}$, where $\delta_{l,m} = 1$ indicates that the m -th high-impact parameter ratio option is selected. Let $\Delta_{l,m} \in \mathbb{R}^{|\theta_{FP}|}$ denote the difference between the quantized model weight θ_Q and the full-precision model weight θ_{FP} when the top m -th percentile of weights in layer l are quantized at a higher bit-width, and the remaining weights in the same layer are quantized at a lower bit-width, $|\cdot|$ signifies the cardinality of a given set. The ratio decision vector for the entire network is defined as $\delta \in \{0, 1\}^{|\mathbb{B}|L}$, where $\delta = \text{concatenate}(\{\delta_l\}_{l=1}^L)$. The corresponding model weight changes induced by the hybrid quantization scheme are captured by:

$$\Delta = \mathbf{D}^\top \delta, \quad (5)$$

where $\mathbf{D} = \text{concatenate}(\{\Delta_{l,m}^\top\}_{l,m}) \in \mathbb{R}^{L|\mathbb{B}| \times |\theta_{FP}|}$ is a matrix in which each row $\mathbf{D}_{l|\mathbb{B}|+m}$ represents the weight changes in the l -th layer under the m -th high-impact parameter ratio option. Substituting Eq. (5) into the objective of Eq. (2) eliminates Δ in favor of the binary selection vector δ . By combining Eqs. (2) and (5), we can transform the objective in Eq. (2) to:

$$1/2 \Delta^\top \mathbf{H} \Delta = 1/2 \delta^\top \mathbf{D} \mathbf{H} \mathbf{D}^\top \delta = 1/2 \delta^\top \mathbf{M} \delta, \quad (6)$$

where the matrix $\mathbf{M} = \mathbf{D} \mathbf{H} \mathbf{D}^\top$. In order to identify the optimal layer-wise ratio, we need to approximate the matrix \mathbf{M} . By replacing Δ in Eq. (2) with $\Delta_{l,m}$, we can approximate each diagonal element

of \mathbf{M} as:

$$\begin{aligned} \mathbf{M}_{|\mathbb{B}|l+m, |\mathbb{B}|l+m} &\approx \mathbf{D}_{l|\mathbb{B}|+m} \mathbf{H} \mathbf{D}_{l|\mathbb{B}|+m}^\top \\ &= \Delta_{l,m}^\top \mathbf{H} \Delta_{l,m} \\ &= 2 \cdot (\mathcal{L}(\theta_{FP} + \Delta_{l,m}, X^{(T)}) \\ &\quad - \mathcal{L}(\theta_{FP}, X^{(T)})). \end{aligned} \quad (7)$$

According to existing works (Li et al., 2021; Ding et al., 2025) that inter-layer interactions are concentrated within individual blocks, we can approximate $\mathbf{H}_{i,j} \approx 0$ if $\theta_{FP,i}$ and $\theta_{FP,j}$ belong to different blocks. Therefore, we have $M_{|\mathbb{B}|l_1+m_1, |\mathbb{B}|l_2+m_2} = \Delta_{l_1,m_1}^\top \mathbf{H} \Delta_{l_2,m_2} = 0$, for any pair of layers l_1 and l_2 that reside in distinct blocks. For layer pairs within the same block, we can approximate $\mathbf{M}_{|\mathbb{B}|l_1+m_1, |\mathbb{B}|l_2+m_2}$ with:

$$\begin{aligned} \mathbf{M}_{|\mathbb{B}|l_1+m_1, |\mathbb{B}|l_2+m_2} &\approx \mathbf{D}_{l_1|\mathbb{B}|+m_1} \mathbf{H} \mathbf{D}_{l_2|\mathbb{B}|+m_2}^\top = \Delta_{l_1,m_1}^\top \mathbf{H} \Delta_{l_2,m_2} \\ &= 1/2 ((\Delta_{l_1,m_1} + \Delta_{l_2,m_2})^\top \mathbf{H} (\Delta_{l_1,m_1} + \Delta_{l_2,m_2}) \\ &\quad - \Delta_{l_1,m_1}^\top \mathbf{H} \Delta_{l_1,m_1} - \Delta_{l_2,m_2}^\top \mathbf{H} \Delta_{l_2,m_2}) \\ &= \mathcal{L}(\theta_{FP} + \Delta_{l_1,m_1} + \Delta_{l_2,m_2}, X^{(T)}) \\ &\quad + \mathcal{L}(\theta_{FP}, X^{(T)}) - \mathcal{L}(\theta_{FP} + \Delta_{l_1,m_1}, X^{(T)}) \\ &\quad - \mathcal{L}(\theta_{FP} + \Delta_{l_2,m_2}, X^{(T)}). \end{aligned} \quad (8)$$

After approximating the elements of matrix \mathbf{M} , we optimize the ratio vector δ by minimizing the quantization error defined in Eq. (2). Let $\mathcal{B}, \mathcal{W} \in \mathbb{R}^{|\mathbb{B}|L}$ be vectors defined by $\mathcal{B}_{i|\mathbb{B}|+j} = \mathbb{B}_j$ and $\mathcal{W}_{i|\mathbb{B}|+j} = |\theta_{FP}^{(j)}|$ for all $0 \leq i < L$ and $0 \leq j < |\mathbb{B}|$, where $|\theta_{FP}^{(j)}|$ denotes the number of parameters in the j -th layer of the full-precision model. The overall optimization is defined as:

$$\begin{aligned} \delta &= \underset{\delta}{\text{argmin}} \delta^\top \mathbf{M} \delta \\ \text{s.t.: } &b^H \delta^\top (\mathcal{B} \odot \mathcal{W}) \\ &\quad + b^N \delta^\top ((1 - \mathcal{B}) \odot \mathcal{W}) \leq C_{\text{target}}, \quad (9) \\ &\sum_{m=1}^{|\mathbb{B}|} \delta_{l,m} = 1, \quad \forall l \in [1, L], \\ &\delta_l \in \{0, 1\}^{|\mathbb{B}|}, \quad \forall l \in [1, L]. \end{aligned}$$

The first constraint encodes the target memory budget: the total bits used by high-impact parameters (at b^H) plus normal parameters (at b^N) must not exceed C_{target} . The second and third constraints enforce that exactly one ratio option is selected for each layer.

Details of the algorithm are provided in Algorithm 1.

Algorithm 1 High-impact parameters ratio optimization

```

1: procedure OPTIMIZE HIGH-IMPACT PARAMETER RATIO( $X^{(T)}$ ,  $\theta_{FP}$ ,  $L$ ,  $B$ ,  $\text{block}(l)$ )
2:    $\triangleright X^{(T)}$ : Calibration data ◁
3:    $\triangleright \theta_{FP}$ : Full-precision model ◁
4:    $\triangleright L$ : Number of layers ◁
5:    $\triangleright B$ : Number of high-impact parameter ratio options ◁
6:    $\triangleright \text{block}(l)$ : Function mapping each layer  $l$  to its block index ◁
7:   initialize the matrix  $\mathbf{M}$ 
8:   for  $l = 1$  to  $L$  do ◁ calculate diagonal elements
9:     for  $i = 1$  to  $B$  do
10:       $\mathbf{M}_{|\mathbb{B}|l+i, |\mathbb{B}|l+i} \leftarrow 2 \cdot (\mathcal{L}(\theta_{FP} + \Delta_{l,i}, X^{(T)}) - \mathcal{L}(\theta_{FP}, X^{(T)}))$  ◁ Eq. (7)
11:     for  $l_1 = 1$  to  $L$  do ◁ calculate non-diagonal elements
12:       for  $m_1 = 1$  to  $B$  do
13:         for  $l_2 = 1$  to  $L$  do
14:           for  $m_2 = 1$  to  $B$  do
15:             if  $\text{block}(l_1) \neq \text{block}(l_2)$  then
16:                $\mathbf{M}_{|\mathbb{B}|l_1+m_1, |\mathbb{B}|l_2+m_2} \leftarrow 0$ 
17:             else
18:                $\mathbf{M}_{|\mathbb{B}|l_1+m_1, |\mathbb{B}|l_2+m_2} \leftarrow \mathcal{L}(\theta_{FP} + \Delta_{l_1, m_1} + \Delta_{l_2, m_2}, X^{(T)})$ 
                  $+ \mathcal{L}(\theta_{FP}, X^{(T)})$ 
                  $- \mathcal{L}(\theta_{FP} + \Delta_{l_1, m_1}, X^{(T)})$ 
                  $- \mathcal{L}(\theta_{FP} + \Delta_{l_2, m_2}, X^{(T)})$ 
19:    $\delta \leftarrow \text{argmin}_{\delta} \delta^{\top} \mathbf{M} \delta$  ◁ optimize for the ratio vector in Eq. (9)
20:   return  $\delta$ 

```

3.3 Hybrid quantization strategy

Based on the optimal ratio for high-impact parameters determined by our quadratic optimization approach, we divide the parameters in each layer into two groups: high-impact parameters and normal parameters with lower impact and apply different quantization strategies to these two groups.

Given $\theta_Q = \{W_l\}_{l=1}^L$, and for each layer $W_l = \{W_l^H, W_l^N\}$ as the set of high-impact and normal parameters for each layer, respectively. The quantization loss of that layer therefore is defined as:

$$\begin{aligned}
 \mathcal{L}(\text{Quant}(W_l), X^{(T)}) & \\
 &= \mathcal{L}(\text{Quant}_H(W_l^H), X^{(T)}) \quad (10) \\
 &+ \mathcal{L}(\text{Quant}_N(W_l^N), X^{(T)})
 \end{aligned}$$

where $\text{Quant}_H(\cdot)$ and $\text{Quant}_N(\cdot)$ respectively denotes the quantizer for the high impact and normal hidden dimensions. For the high impact dimensions, we adopt the AdaRound (Nagel et al., 2020) quantization approach that adapts to data and task loss, which is defined as:

$$\begin{aligned}
 \text{Quant}_H(W_l^H) &= s \cdot \text{Clip}\left(\lfloor W_l^H/s \rfloor + V_l^H, \right. \\
 &\quad \left. 0, 2^{b^H} - 1\right) \quad (11) \\
 \text{s.t.: } V_l^H &\in [0, 1],
 \end{aligned}$$

where s is scale, and V_l^H is the learnable weight-rounding parameters with same dimension as W_l^H .

During the optimization (Eq. (10)), the elements of the rounding function V_l^H are encouraged toward binary values (0 or 1) using a regularization loss term added to the main objective in Eq. (10):

$$\mathcal{L}_{\text{reg}} = (1 - |2V_l^H - 1|^\gamma), \quad (12)$$

where γ is an annealing factor that starts high and decreases during optimization to encourage convergence to binary values (Nagel et al., 2020).

For the normal parameters, we adopt learnable weight clipping (Shao et al., 2024) as follows:

$$\begin{aligned}
 \text{Quant}_N(W_l^N) &= \text{Clamp}(\lfloor W_l^N/h \rfloor, 0, 2^{b^N} - 1), \\
 h &= \frac{\alpha \max(W_l^N) - \beta \min(W_l^N)}{2^{b^N} - 1}, \quad (13)
 \end{aligned}$$

where α and β are learnable clipping values for the upper and lower bounds of the weight values.

This hybrid approach efficiently allocates computational resources by prioritizing high-impact parameters, greatly reducing overhead while preserving overall model performance.

4 Experiments

4.1 Experimental setup

Calibration dataset and evaluation metrics. For calibration, we randomly sample 128 sequences, each containing 2048 tokens, from the WikiText-2 dataset (Merity et al., 2016). For evaluation metrics, we report perplexity (lower is better) specifically on WikiText-2 (Merity et al., 2016) and C4 (Raffel et al., 2020) for language modeling tasks. Following current SOTA works (Shao et al., 2024; Kim et al., 2024), we also evaluate in the zero-shot tasks, including Commonsense reasoning: HellaSwag (Zellers et al., 2019), PIQA (Bisk et al., 2020), WinoGrande (Sakaguchi et al., 2021), and World knowledge: ARC-easy/challenge (Clark et al., 2018). We compare our method against several state-of-the-art PTQ methods: GPTQ (Frantar et al., 2023), AWQ (Lin et al., 2024), SqueezeLLM (Kim et al., 2024), CBQ (Ding et al., 2025), and OmniQuant (Shao et al., 2024).

Implementation details. We evaluate our method on several widely-used LLMs, including LLaMA-2-7B and LLaMA-2-13B (Touvron et al., 2023). Following OmniQuant (Shao et al., 2024), we employ a block reconstruction loss function and validate both per-group and per-channel weight quantization. The notation $W2A16g64$ denotes the quantization with 2-bit per-group weight-only, FP16 activation and a group size of 64.¹ All experiments were conducted on NVIDIA A100 GPUs. We set the candidate ratio set $\mathbb{B} = \{0.02, 0.05, 0.1, 0.15, 0.2\}$ when optimizing for the high-impact parameter ratio δ in Eq. (9). For 2-bit quantization, we use a higher-precision bit-width of 3 bits. Similarly, for 3-bit quantization, we set the higher-precision bit-width to 4 bits. For quantizing the high-impact parameters, we optimize the learnable rounding parameter for 5,000 iterations. The optimizer used is Adam with a learning rate of 10^{-3} and an L2 regularization of 10^{-5} . The annealing factor γ in Eq. (12) is

¹Throughout the paper, $WxA16$ denotes *weight-only* quantization at x bits with activations kept at FP16, following the convention of (Shao et al., 2024; Frantar et al., 2023; Lin et al., 2024; Kim et al., 2024).

initially set to 20 and gradually decreased to 2 during optimization. For the remaining parameters, we adopt the learnable weight clipping method from OmniQuant (Shao et al., 2024).

4.2 Experimental results

Evaluation on generation datasets with perplexity. Results in Tables 1 and 2 show the performance of our method in text generation on C4, WikiText-2 using weight-only quantized LLaMA models. Our method consistently outperforms existing methods like AWQ and OmniQuant (Shao et al., 2024), particularly at the low-bit configurations, such as 2-bit and 3-bit, on the LLaMA-2-7B model. Specifically, in the W2A16 setting (Table 1), OmniQuant only achieves a perplexity of 37.37 while our method substantially drives this result to a perplexity of 9.40 on WikiText-2 dataset. Our method also improves the perplexity in the W2A16g64 setting with a gap of 0.24 and 0.48 compared to CBQ (Ding et al., 2025) on WikiText-2 and C4 dataset, respectively. This indicates that extremely low-bit weight quantization requires a careful adjustment for each parameter. In the W3A16 setting (Table 2), the proposed approach improves OmniQuant by 0.25 and 0.36 perplexity scores on WikiText-2 and C4 datasets, respectively.

Evaluation on downstream tasks. Table 3 shows the evaluation of our method in multiple zero-shot benchmarks for LLaMA-2-7B at various quantization settings (i.e., W2A16g128 and W3A16g128). In the W2A16g128 setting, the proposed method achieves an average accuracy of 49.04%, outperforming OmniQuant by 1.45% and GPTQ by 10.51%. In the W3A16g128 setting, our method achieves an average accuracy of 60.76%, significantly surpassing GPTQ that performs at 52.39% accuracy. In W3A16g128 quantization, the proposed method reduces the accuracy gap between the quantized and full-precision models to approximately 4%, demonstrating the effectiveness of our approach in preserving model performance even under low-bit quantization.

4.3 Ablation studies

The effectiveness of determining the ratio of high-impact parameters. We analyze the impact of our approach for determining the ratio of high-impact parameters on model performance. We conduct experiments on LLaMA-2-7B with W2A16 quantization using two settings: (1) a fixed ratio of high-impact parameters across all layers, and (2)

Model	Method	Avg. bit	W2A16		Avg. bit	W2A16g128		Avg. bit	W2A16g64	
			C4	Wiki2		C4	Wiki2		C4	Wiki2
LLaMA-2-7B	FP16	16	6.97	5.47	16	6.97	5.47	16	6.97	5.47
	GPTQ	2.00	–	–	2.15	33.70	36.77	2.30	19.40	20.85
	OmniQuant	2.00	90.64	37.37	2.15	15.02	11.06	2.24	12.72	9.62
	CBQ	2.00	–	–	–	–	–	2.24	11.30	8.01
	Ours	2.20	14.64	9.40	2.25	11.46	8.13	2.30	10.82	7.77
LLaMA-2-13B	FP16	16	6.47	4.88	16	6.47	4.88	16	6.47	4.88
	GPTQ	2.00	–	–	2.15	20.97	28.14	2.30	12.48	22.44
	OmniQuant	2.00	26.76	17.21	2.15	11.05	8.26	2.24	10.05	7.56
	Ours	2.20	14.38	9.86	2.25	10.06	7.41	2.30	9.37	6.96

Table 1: Perplexity scores (\downarrow) on WikiText-2 and C4 for 2-bit weight-only quantization on LLaMA-2 models. GPTQ and OmniQuant results are from (Shao et al., 2024); CBQ results are from (Ding et al., 2025). **Bold** indicates the best result in each column.

Model	Method	Avg. bit	W3A16		Avg. bit	W3A16g128		Avg. bit	W3A16g64	
			C4	Wiki2		C4	Wiki2		C4	Wiki2
LLaMA-2-7B	FP16	16	6.97	5.47	16	6.97	5.47	16	6.97	5.47
	GPTQ	3.00	9.81	8.37	3.15	8.28	6.74	3.30	8.20	6.62
	AWQ	3.00	17.19	15.39	3.15	7.84	6.24	3.24	7.98	6.07
	OmniQuant	3.00	8.65	6.58	3.15	7.75	6.03	3.24	7.63	5.97
	SqueezeLLM	3.13	7.71	6.16	3.24	7.39	5.89	3.24	7.51	5.96
	CBQ	3.00	–	–	–	–	–	3.24	7.56	5.89
	Ours	3.13	7.84	6.03	3.25	7.55	5.85	3.30	7.49	5.82
LLaMA-2-13B	FP16	16	6.47	4.88	16	6.47	4.88	16	6.47	4.88
	GPTQ	3.00	8.02	6.44	3.15	7.24	5.63	3.30	7.10	5.56
	AWQ	3.00	8.91	6.45	3.15	6.94	5.32	3.24	7.13	5.25
	OmniQuant	3.00	7.44	5.58	3.15	6.98	5.28	3.24	6.91	5.24
	SqueezeLLM	3.13	6.96	5.39	3.24	6.82	5.23	3.24	6.83	5.19
	Ours	3.13	7.08	5.32	3.25	6.78	5.22	3.30	6.80	5.16

Table 2: Perplexity (\downarrow) of 3-bit weight-only quantization on LLaMA-2 models. OmniQuant results and AWQ (W3A16g128) results are from (Shao et al., 2024), while SqueezeLLM (W3A16g64) results for LLaMA-2-7B are from (Kim et al., 2024). **Bold** indicates the best result in each column.

layer-specific ratios determined by our approach. In both settings, we quantize high-impact parameters and remaining parameters using OmniQuant. As shown in Table 4, the layer-specific ratios determined by our approach improves performance by 2.4% on average across downstream tasks and decreases perplexity by 0.15 compared to the fixed-ratio approach. These results demonstrate the effectiveness of the proposed quadratic optimization framework in determining the optimal ratio of high-impact parameters across layers.

The effectiveness of the hybrid quantization strategy. To validate the impact of the hybrid quantization strategy on the quantized model performance, we conduct ablation studies on the LLaMA-2-7B with W2A16 quantization. As

shown in Table 4, the hybrid quantization strategy improves the performance of LLMs when used with either fixed or optimized ratios of high-impact parameters. Specifically, when combining the fixed ratio of high-impact parameters with the hybrid quantization strategy, the performance of LLMs improves by 1.06% on average across downstream tasks, while enhancing perplexity by 0.46 on average compared to using only the fixed ratio of high-impact parameters. Moreover, the performance of quantized models further improves when combining the optimized ratio of high-impact parameters with the hybrid quantization strategy. These improvements are 2.76% on average across downstream tasks and 1.22 perplexity compared to using only the fixed ratio of high-impact parameters. These results indicate the effectiveness of the hy-

Bitwidths	Methods	PiQA	ArcE	ArcC	HellaSwag	WinoGrande	Avg.
FP16	-	78.07	76.34	43.51	57.17	69.21	64.87
W2A16 g128	GPTQ	58.21	33.75	19.79	29.60	51.30	38.53
	OmniQuant	64.79	51.13	24.83	40.30	56.90	47.59
	Ours	68.82	51.30	26.19	41.65	57.22	49.04
W3A16 g128	GPTQ	76.65	73.69	40.52	54.43	66.61	52.39
	OmniQuant	77.37	68.01	37.20	54.21	66.30	60.62
	Ours	77.80	68.51	37.37	54.33	66.32	60.87

Table 3: The accuracy of 5 common sense reasoning tasks (\uparrow) on LLaMa-2 models. **Bold** indicates the best result in each column.

Methods	PiQA	ArcE	ArcC	HellaSwag	WinoGrande	Avg. (\uparrow)	PPL (\downarrow)
FP16	78.07	76.34	43.51	57.17	69.21	64.87	5.68
OmniQuant	56.20	32.78	22.44	29.36	50.91	38.34	37.27
+ uniform ratio	64.25	48.61	26.10	38.67	52.33	45.99	10.66
+ optimal ratio	67.08	52.27	27.56	41.03	53.42	48.27	9.80
+ uniform ratio + hybrid quantization	66.38	49.24	26.51	39.73	53.38	47.05	10.2
+ optimal ratio + hybrid quantization (Ours)	67.14	53.89	28.10	41.13	53.49	48.75	9.40

Table 4: Ablation study on LLaMA-2-7B with W2A16 quantization. We report WikiText-2 perplexity and average accuracy across downstream tasks.

brid quantization strategy.

Complexity. Regarding time complexity, our framework is designed for efficiency by limiting inter-layer dependency analysis to layers within the same block (see Algorithm 1). For a network with \mathcal{B} blocks, each containing L_b layers, the method performs only $\frac{L_b(L_b-1)}{2}$ validation evaluations per block. As a result, the overall time complexity is $O(\mathcal{B}L_b(L_b-1))$, which scales linearly with the number of blocks. This block-wise design ensures that the method remains highly computationally efficient and practical during the quantization process, particularly for large-scale networks composed of many small blocks.

5 Conclusion

In this paper, we addressed the post-training quantization for large language models, with a particular focus on the critical issue of high-impact parameters that significantly influence quantization performance. Specifically, we proposed a quadratic optimization framework that determines the optimal layer-specific ratios of these high-impact parameters. This approach enables more accurate allocation of quantization precision based on each layer’s contribution to the overall model performance. Furthermore, the proposed approach allows us to apply a hybrid quantization strategy for post-training quantization on LLMs, which lever-

ages advanced quantization methods exclusively for high-impact parameters, while employing computationally efficient methods to remaining parameters. This strategy achieves an effective trade-off between computational efficiency and model accuracy. Extensive experimental results across various model sizes and quantization configurations demonstrate that our approach outperforms state-of-the-art methods, particularly in extremely low bit-width scenarios.

Limitation

While the proposed method improves performance over existing approaches, it relies on manually assigning predetermined high bit-widths to parameters identified as high-impact. This manual configuration may not yield optimal results under a given resource constraint. Future work could focus on dynamically setting the bit-width for these high-impact parameters during training, enabling more efficient and adaptive quantization strategies.

References

- Saleh Ashkboos, Amirkeivan Mohtashami, Maximilian Croci, Bo Li, Pashmina Cameron, Martin Jaggi, Dan Alistarh, Torsten Hoefer, and James Hensman. 2024. Quarot: Outlier-free 4-bit inference in rotated llms. In *NeurIPS*.
- Yonatan Bisk, Rowan Zellers, Jianfeng Gao, Yejin Choi,

- and 1 others. 2020. Piqa: Reasoning about physical commonsense in natural language. In *AAAI*.
- Weihan Chen, Peisong Wang, and Jian Cheng. 2021. Towards mixed-precision quantization of neural networks via constrained optimization. In *ICCV*.
- Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv*.
- Wanyun Cui and Qianle Wang. 2024. Cherry on top: Parameter heterogeneity and quantization in large language models. In *NeurIPS*.
- Zihao Deng, Xin Wang, Sayeh Sharify, and Michael Orshansky. 2023. Mixed-precision quantization with cross-layer dependencies. *arXiv preprint arXiv:2307.05657*.
- Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. 2022. Gpt3. int8 (): 8-bit matrix multiplication for transformers at scale. In *NeurIPS*.
- Xin Ding, Xiaoyu Liu, Zhijun Tu, Yun Zhang, Wei Li, Jie Hu, Hanting Chen, Yehui Tang, Zhiwei Xiong, Baoqun Yin, and Yunhe Wang. 2025. CBQ: Cross-block quantization for large language models. In *ICLR*.
- Zhen Dong, Zhewei Yao, Amir Gholami, Michael W. Mahoney, and Kurt Keutzer. 2019. Hawq: Hessian aware quantization of neural networks with mixed-precision. In *ICCV*.
- Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. 2023. GPTQ: Accurate post-training quantization for generative pre-trained transformers. In *ICLR*.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *ICLR*.
- Yushi Huang, Ruihao Gong, Jing Liu, Tianlong Chen, and Xianglong Liu. 2024. TFMQ-DM: Temporal feature maintenance quantization for diffusion models. In *CVPR*.
- Yongkweon Jeon, Chungman Lee, and Ho-young Kim. 2023. Genie: Show me the data for quantization. In *CVPR*.
- Sehoon Kim, Coleman Hooper, Amir Gholami, Zhen Dong, Xiuyu Li, Sheng Shen, Michael W Mahoney, and Kurt Keutzer. 2024. SqueezeLLM: Dense-and-sparse quantization. In *ICML*.
- Changhun Lee, Jungyu Jin, Taesu Kim, Hyungjun Kim, and Eunhyeok Park. 2024. Owq: Outlier-aware weight quantization for efficient fine-tuning and inference of large language models. In *AAAI*.
- Xiuyu Li, Yijiang Liu, Long Lian, Huanrui Yang, Zhen Dong, Daniel Kang, Shanghang Zhang, and Kurt Keutzer. 2023. Q-Diffusion: Quantizing diffusion models. In *ICCV*.
- Yuhang Li, Ruihao Gong, Xu Tan, Yang Yang, Peng Hu, Qi Zhang, Fengwei Yu, Wei Wang, and Shi Gu. 2021. BRECCQ: Pushing the limit of post-training quantization by block reconstruction. In *ICLR*.
- Chen Lin, Bo Peng, Zheyang Li, Wenming Tan, Ye Ren, Jun Xiao, and Shiliang Pu. 2023. Bit-shrinking: Limiting instantaneous sharpness for improving post-training quantization. In *CVPR*.
- Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Weiming Chen, Wei-Chen Wang, Guangxuan Xiao, Xingyu Dang, Chuang Gan, and Song Han. 2024. Awq: Activation-aware weight quantization for on-device llm compression and acceleration. In *MLSys*.
- Jiawei Liu, Lin Niu, Zhihang Yuan, Dawei Yang, Xinggang Wang, and Wenyu Liu. 2023a. Pd-quant: Post-training quantization based on prediction difference metric. In *CVPR*.
- Jing Liu, Ruihao Gong, Xiuying Wei, Zhiwei Dong, Jianfei Cai, and Bohan Zhuang. 2023b. Qllm: Accurate and efficient low-bitwidth quantization for large language models. *arXiv*.
- Zechun Liu, Changsheng Zhao, Igor Fedorov, Bilge Soran, Dhruv Choudhary, Raghuraman Krishnamoorthi, Vikas Chandra, Yuandong Tian, and Tijmen Blankevoort. 2025. Spinqant: LLM quantization with learned rotations. In *ICLR*.
- Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2016. Pointer sentinel mixture models. *arXiv preprint arXiv:1609.07843*.
- Markus Nagel, Rana Ali Amjad, Mart Van Baalen, Christos Louizos, and Tijmen Blankevoort. 2020. Up or down? adaptive rounding for post-training quantization. In *ICML*.
- Cuong Pham, Anh Dung Hoang, Cuong C Nguyen, Trung Le, Dinh Phung, Gustavo Carneiro, and Thanh-Toan Do. 2024. MetaAug: Meta-data augmentation for post-training quantization. In *ECCV*.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *JMLR*.
- Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavathula, and Yejin Choi. 2021. Winogrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, 64(9):99–106.
- Yuzhang Shang, Zhihang Yuan, Bin Xie, Bingzhe Wu, and Yan Yan. 2023. Post-training quantization on diffusion models. In *CVPR*.

- Wenqi Shao, Mengzhao Chen, Zhaoyang Zhang, Peng Xu, Lirui Zhao, Zhiqian Li, Kaipeng Zhang, Peng Gao, Yu Qiao, and Ping Luo. 2024. Omniquant: Omnidirectionally calibrated quantization for large language models. In *ICLR*.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and 1 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv*.
- Albert Tseng, Jerry Chee, Qingyao Sun, Volodymyr Kuleshov, and Christopher De Sa. 2024. Quip#: Even better llm quantization with hadamard incoherence and lattice codebooks. In *ICML*.
- Kuan Wang, Zhijian Liu, Yujun Lin, Ji Lin, and Song Han. 2019. Haq: Hardware-aware automated quantization with mixed precision. In *CVPR*.
- Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022a. Emergent abilities of large language models. *TMLR*.
- Xiuying Wei, Ruihao Gong, Yuhang Li, Xianglong Liu, and Fengwei Yu. 2022b. Qdrop: randomly dropping quantization for extremely low-bit post-training quantization. In *ICLR*.
- Xiuying Wei, Yunchen Zhang, Yuhang Li, Xiangguo Zhang, Ruihao Gong, Jinyang Guo, and Xianglong Liu. 2023. Outlier suppression+: Accurate quantization of large language models by equivalent and optimal shifting and scaling. *arXiv preprint arXiv:2304.09145*.
- Xiuying Wei, Yunchen Zhang, Xiangguo Zhang, Ruihao Gong, Shanghang Zhang, Qi Zhang, Fengwei Yu, and Xianglong Liu. 2022c. Outlier suppression: Pushing the limit of low-bit transformer language models. In *NeurIPS*.
- Guangxuan Xiao, Ji Lin, Mickael Seznec, Hao Wu, Julien Demouth, and Song Han. 2023. Smoothquant: Accurate and efficient post-training quantization for large language models. In *ICML*.
- Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a machine really finish your sentence? In *ACL*.
- Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, and 1 others. 2022. Opt: Open pre-trained transformer language models. *arXiv*.

A Appendix

A.1 More Experimental Results and Ablation Studies

Candidate ratio set \mathbb{B}	Wiki2 (\downarrow)	C4 (\downarrow)
{0.05, 0.10, 0.15}	57.18	59.66
{0.02, 0.05, 0.10, 0.15, 0.20}	56.91	58.79

Table A.1: Sensitivity to the candidate ratio set \mathbb{B} on OPT-125M with W2A16g128. Finer-grained sets yield slightly better perplexity at the cost of more M evaluations ($|\mathbb{M}|$ scales as $|\mathbb{B}|^2$).

Setting	Method	Wiki2 (\downarrow)	C4 (\downarrow)
W2A16g128	OmniQuant	75.43	80.01
	Ours	56.91	58.79
W2A16g64	OmniQuant	62.56	64.01
	Ours	48.45	49.61

Table A.2: Perplexity results on OPT-125M at 2-bit weight-only quantization.

Calibration Dataset	W2A16g128		W3A16g128	
	WikiText2	C4	WikiText2	C4
WikiText2	8.13	11.46	7.55	5.85
C4	8.25	11.23	7.62	5.80

Table A.3: Perplexity scores of LLaMA-2-7B models using different calibration datasets.

Sensitivity to the candidate ratio set \mathbb{B} . We vary the candidate ratio set \mathbb{B} on OPT-125M with W2A16g128 quantization. As shown in Table A.1, finer-grained candidate sets yield slightly better perplexity, at the cost of more matrix-M evaluations, since $|\mathbb{M}|$ grows as $|\mathbb{B}|^2$. In practice, the default set {0.02, 0.05, 0.10, 0.15, 0.20} provides a good trade-off.

Results on OPT-125M. We also evaluate on OPT-125M with W2A16g128 and W2A16g64 quantization settings. As shown in Table A.2, the proposed method substantially outperforms OmniQuant in the extreme 2-bit setting on OPT-125M.

Ablation study using different calibration datasets. We perform experiments using different calibration datasets, including WikiText2 and C4, each consisting of 128 samples with 2048 tokens. As shown in Table A.3, the variance in perplexity scores when quantized with WikiText2 and C4 is small.

Ablation study on the choice of bit-width for high-impact parameters given the same resource constraint. Given the same resource constraint, we compare the performance of our approach with different bit-widths for high-impact parameters. As shown in Table A.4, given the same resource constraint, when quantizing the model to 2-bit, setting the bit-width for high-impact parameters to 3 achieves better performance than setting it to 4. When quantizing the model to 3-bit, setting the bit-width for high-impact parameters to 4 achieves better performance than setting it to 16.

A.2 More visualization

We visualize the distribution of average Fisher information across channel dimensions for selected layers in the LLaMA-2-7B model. As shown in Figure A.1, the attention query projection layer at block 1 and mlp down projection layer at block 5 exhibit channels with significantly higher importance scores compared to others, while the overall ratio of high-impact parameters could be small. In contrast, the key projection layer in block 2 and the value projection layer in block 5 may contain a higher ratio of high-impact parameters than the previously mentioned layers.

Visualization of optimized high-impact parameter ratio δ across layers. Fig. A.2 shows the visualization of the distribution of high-impact parameter ratio δ across layers of the LLaMA-2-7B model in W3A16 quantization setting. In general, the proposed approach allocates various ratios of high-impact parameters to the same layer type across different blocks. Additionally, the values of these ratios are often higher in the self-attention value (v) projection layer compared to the self-attention query (q) projection layer. This indicates that certain layers contain a larger proportion of high-impact parameters, thus may have a higher influence to the performance of the quantized model.

Methods	Bit-width	Higher bit-width b^H	Avg. bit	Wikitext2 (PPL ↓)	C4 (PPL ↓)
OmniQuant		-	2.0	37.37	90.64
Ours	W2A16	16	2.2	15.39	26.93
		4	2.2	11.05	15.14
		3	2.2	9.40	14.64
OmniQuant		-	2.0	11.06	15.02
Ours	W2A16g128	16	2.25	10.05	13.51
		4	2.25	8.90	11.98
		3	2.25	8.13	11.46
OmniQuant		-	2.0	9.62	12.72
Ours	W2A16g64	16	2.3	9.12	11.92
		4	2.3	8.25	11.01
		3	2.3	7.77	10.82
OmniQuant		-	3.0	6.58	8.65
Ours	W3A16	16	3.13	6.26	8.16
		4	3.13	6.03	7.84
OmniQuant		-	3.0	6.03	7.75
Ours	W3A16g128	16	3.25	6.03	7.71
		4	3.25	5.85	7.55
OmniQuant		-	3.0	5.97	7.63
Ours	W3A16g64	16	3.3	5.93	7.57
		4	3.3	5.82	7.49

Table A.4: The perplexity of quantized LLaMa2-7B models for different high-impact parameter bit-widths.

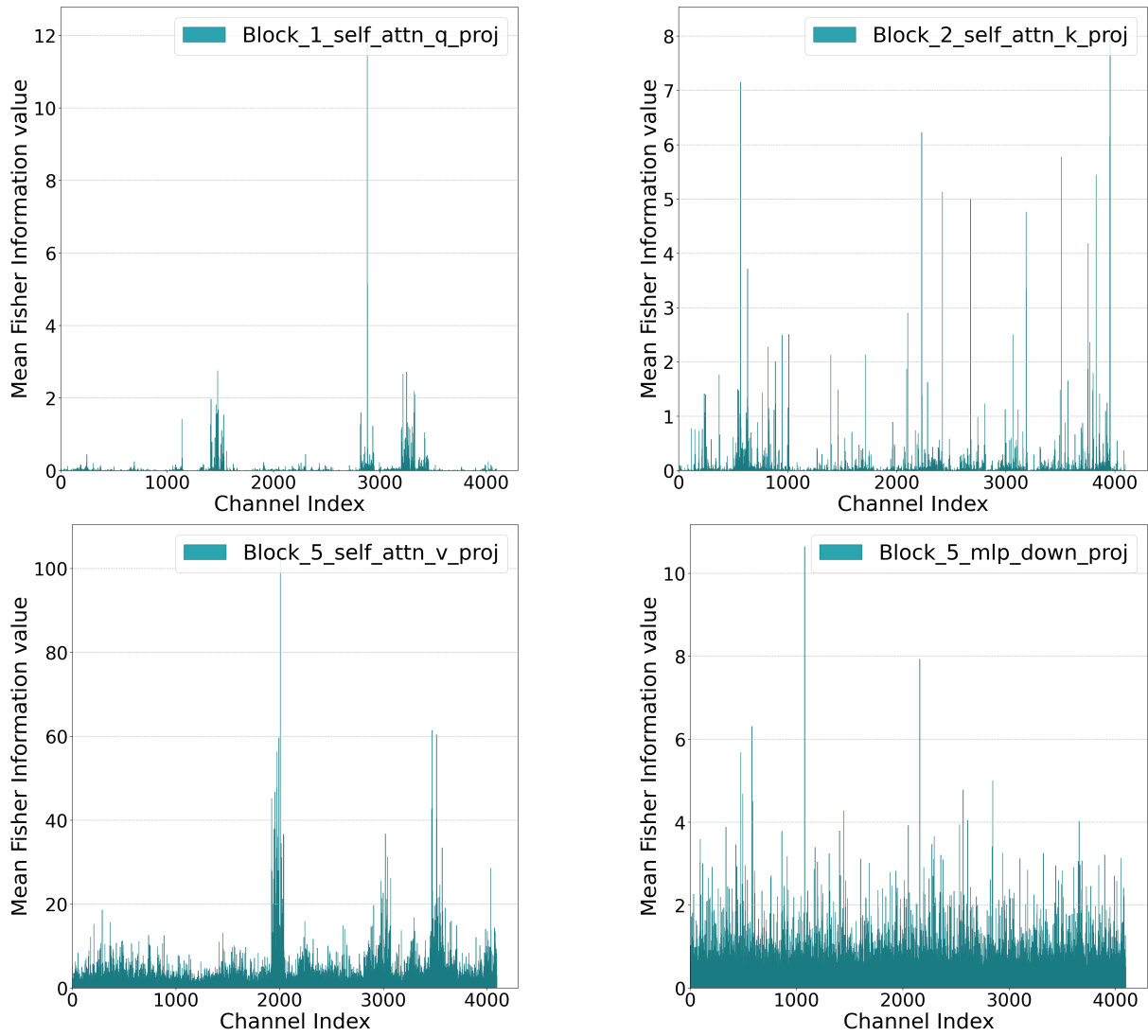


Figure A.1: Distribution of average Fisher information across channel dimensions at several layers in the LLaMA-2-7B model. There is a significant difference in importance scores for each layer, indicating that impactful parameters are concentrated in a subset of channels. Additionally, the ratio of high-impact parameters differs across layers.

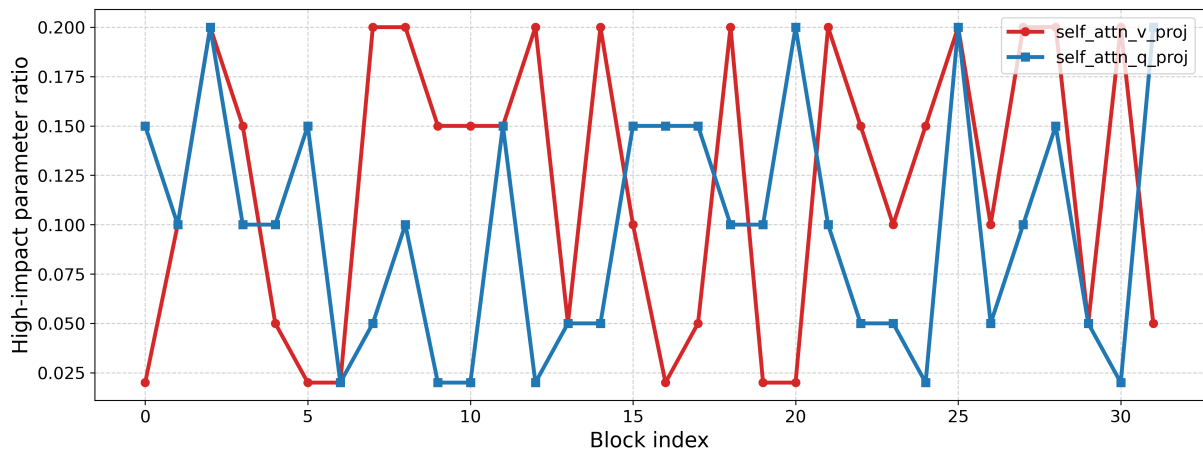


Figure A.2: High-impact parameter ratio for selected layers across blocks in the LLaMA-2-7B model with W3A16 quantization.