GPTQ

Kamil Mardanshin
Foma Shipilov

HSE
Skoltech

2024

Optimal Brain Damage (OBD). Problem statement

We model a layer ℓ as a function $f_\ell(\mathbf{X}_\ell, \mathbf{W}_\ell)$ acting on inputs $\mathbf{X}_\ell$, parametrized by weights $\mathbf{W}_\ell$. For compressed weights $\widehat{\mathbf{W}}_\ell$, a loss $\mathcal{L}$ and a generic compression constraint $\mathcal{C}(\widehat{\mathbf{W}}_\ell) > C$, we try to minimize this objective in post training quantization

$$\text{argmin}_{\widehat{\mathbf{W}}_\ell} \ \mathbb{E}_{\mathbf{X}_\ell} \mathcal{L}(f_\ell(\mathbf{X}_\ell, \mathbf{W}_\ell), f_\ell(\mathbf{X}_\ell, \widehat{\mathbf{W}}_\ell)) \quad \text{s.t.} \quad \mathcal{C}(\widehat{\mathbf{W}}_\ell) > C$$

For $d_{\text{row}} \times d_{\text{col}}$ weights $\mathbf{W}_\ell$ and the input $\mathbf{X}_\ell$ has dimensions $d_{\text{col}} \times N$ we minimize [1] [3]

$$\text{argmin}_{\widehat{\mathbf{W}}_\ell} \ ||\mathbf{W}_\ell \mathbf{X}_\ell - \widehat{\mathbf{W}}_\ell \mathbf{X}_\ell||_2^2 \quad \text{s.t.} \quad \mathcal{C}(\widehat{\mathbf{W}}_\ell) > C. \tag{1}$$

The exact solution is an NP-hard problem.

Optimal Brain Surgeon (OBS). Greedy approach

Taylor expansion of the error of a neural network [2]

$$\delta E = \left(\frac{\partial E}{\partial \mathbf{w}}\right)\delta\mathbf{w} + \frac{1}{2}\delta\mathbf{w}^T\mathbf{H}\delta\mathbf{w} + O\left(\|\delta\mathbf{w}\|^3\right) \tag{2}$$

where $\mathbf{H}$ is the Hessian matrix. For a network trained to a local minimum, the first order term vanishes. Our goal is to set one of the weights (w_p) to zero by adding a small $\delta\mathbf{w}$. A constrained optimization problem arises:

$$\min_p \min_{\delta\mathbf{w}} \frac{1}{2}\delta\mathbf{w}^T\mathbf{H}\delta\mathbf{w}, \text{ such that } \mathbf{e}_p^T\delta\mathbf{w} + w_p = 0 \tag{3}$$

The Lagrangian is

$$L(\delta\mathbf{w}, \lambda) = \frac{1}{2}\delta\mathbf{w}^T\mathbf{H}\delta\mathbf{w} + \lambda(\mathbf{e}_p^T\delta\mathbf{w} + w_p) \tag{4}$$

Optimal Brain Surgeon (OBS). Greedy approach

The derivatives are

$$\nabla_{\delta \mathbf{w}} L = \mathbf{H}\delta\mathbf{w} + \lambda \mathbf{e}_p \tag{5}$$

$$\nabla_{\lambda} L = \mathbf{e}_p^T \delta\mathbf{w} + w_p \tag{6}$$

Solving for a saddle point

$$\delta\mathbf{w} = -\lambda \mathbf{H}^{-1}\mathbf{e}_p$$

$$\delta\mathbf{w}_p = -w_p \;\Rightarrow$$

$$\lambda = w_p / [\mathbf{H}^{-1}]_{pp}$$

$$\delta\mathbf{w} = -\frac{w_p}{[\mathbf{H}^{-1}]_{pp}} \cdot \mathbf{H}^{-1}\mathbf{e}_p, \quad w_p = \mathrm{argmin}_{w_p} \frac{w_p^2}{[\mathbf{H}^{-1}]_{pp}}$$

Optimal Brain Surgeon (OBS). Greedy approach

Likewise, for weight quantization we can write

$$L(\delta \mathbf{w}, \lambda) = \frac{1}{2} \delta \mathbf{w}^T \mathbf{H} \delta \mathbf{w} + \lambda (\mathbf{e}_p^T \delta \mathbf{w} + w_p - \text{quant}(w_p)) \tag{7}$$

so that $\delta \mathbf{w}_p + w_p = \text{quant}(w_p)$ the small addition quantizes the weight instead of zeroing it. The solution is analogous

$$\delta \mathbf{w} = -\frac{w_p - \text{quant}(w_p)}{[\mathbf{H}^{-1}]_{pp}} \cdot \mathbf{H}^{-1} \mathbf{e}_p, \quad w_p = \text{argmin}_{w_p} \frac{(w_p - \text{quant}(w_p))^2}{[\mathbf{H}^{-1}]_{pp}}$$

Optimal Brain Quantizer (OBQ). Hessian update

Lemma (Row & Column Removal)

Given an invertible $d_{col} \times d_{col}$ matrix $\mathbf{H}$ and its inverse $\mathbf{H}^{-1}$, we want to efficiently compute the inverse of $\mathbf{H}$ with row and column p removed, which we denote by $\mathbf{H}_{-p}$. This can be accomplished through the following formula:

$$\mathbf{H}_{-p}^{-1} = \left(\mathbf{H}^{-1} - \frac{1}{[\mathbf{H}^{-1}]_{pp}} \mathbf{H}_{:,p}^{-1} \mathbf{H}_{p,:}^{-1}\right)_{-p}, \tag{8}$$

which corresponds to performing Gaussian elimination of row and column p in $\mathbf{H}^{-1}$ followed by dropping them completely. This has $\Theta(d_{col}^2)$ time complexity.

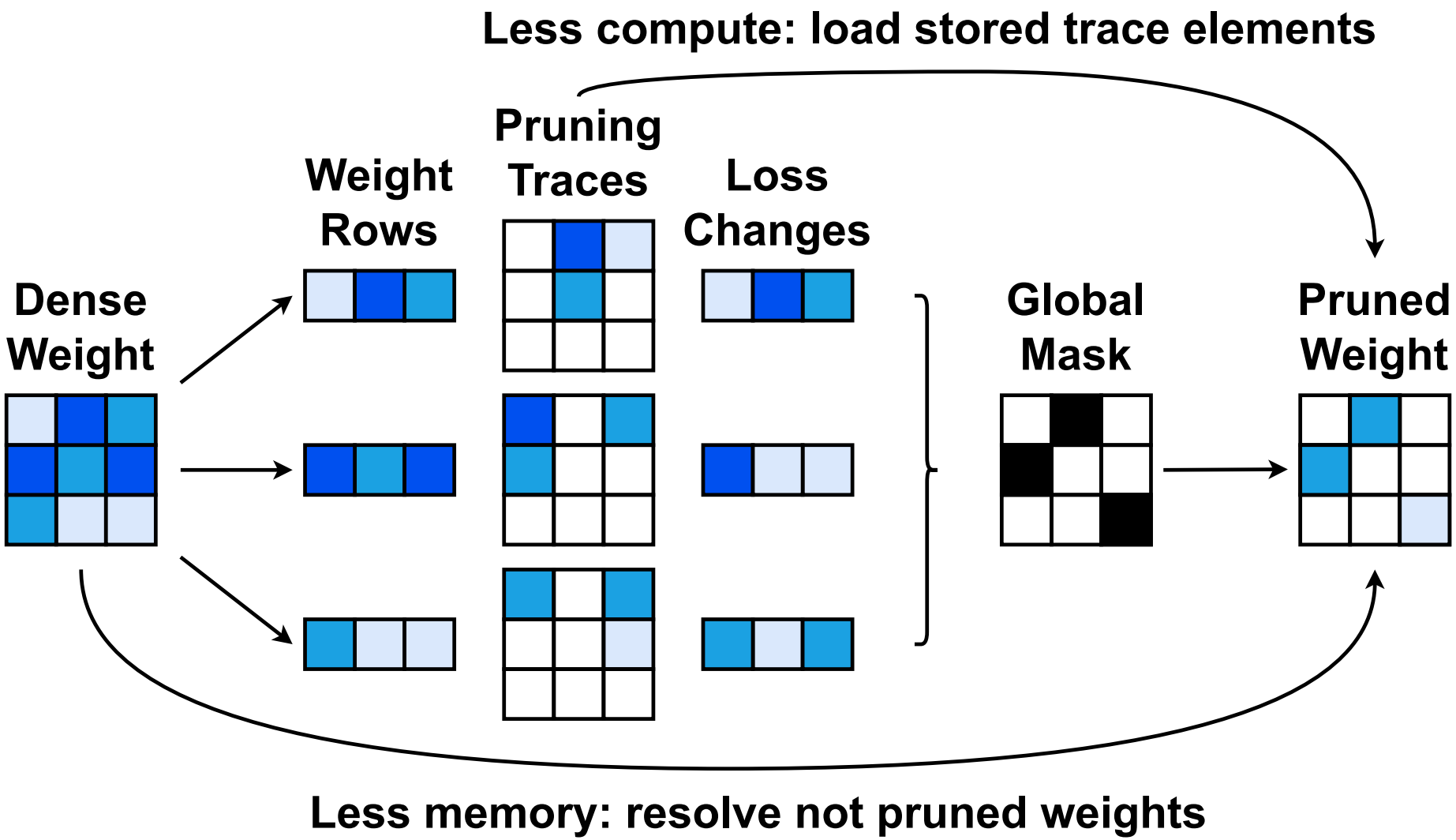
Optimal Brain Quantizer (OBQ). Row implemetation

Algorithm 1 Quantize $k \leq d_{\text{col}}$ weights from row $\mathbf{w}$ with inverse Hessian $\mathbf{H}^{-1} = (2\mathbf{X}\mathbf{X}^\top)^{-1}$ according to OBS in $O(k \cdot d_{\text{col}}^2)$ time.

$M = \{1, \ldots, d_{\text{col}}\}$
for $i = 1, \ldots, k$ **do**
  $p \leftarrow \text{argmin}_{p \in M} \frac{1}{[\mathbf{H}^{-1}]_{pp}} \cdot (\text{quant}(w_p) - w_p)^2$
  $\mathbf{w} \leftarrow \mathbf{w} - \mathbf{H}^{-1}_{:,p} \frac{1}{[\mathbf{H}^{-1}]_{pp}} \cdot (\text{quant}(w_p) - w_p)$
  $\mathbf{H}^{-1} \leftarrow \mathbf{H}^{-1} - \frac{1}{[\mathbf{H}^{-1}]_{pp}} \mathbf{H}^{-1}_{:,p} \mathbf{H}^{-1}_{p,:}$
  $M \leftarrow M - \{p\}$
end for

$O(d_{row} d_{col})$ Hessian updates with time complexity $O(d_{col}^2)$. Global OBQ time complexity $O(d_{row} d_{col}^3)$.

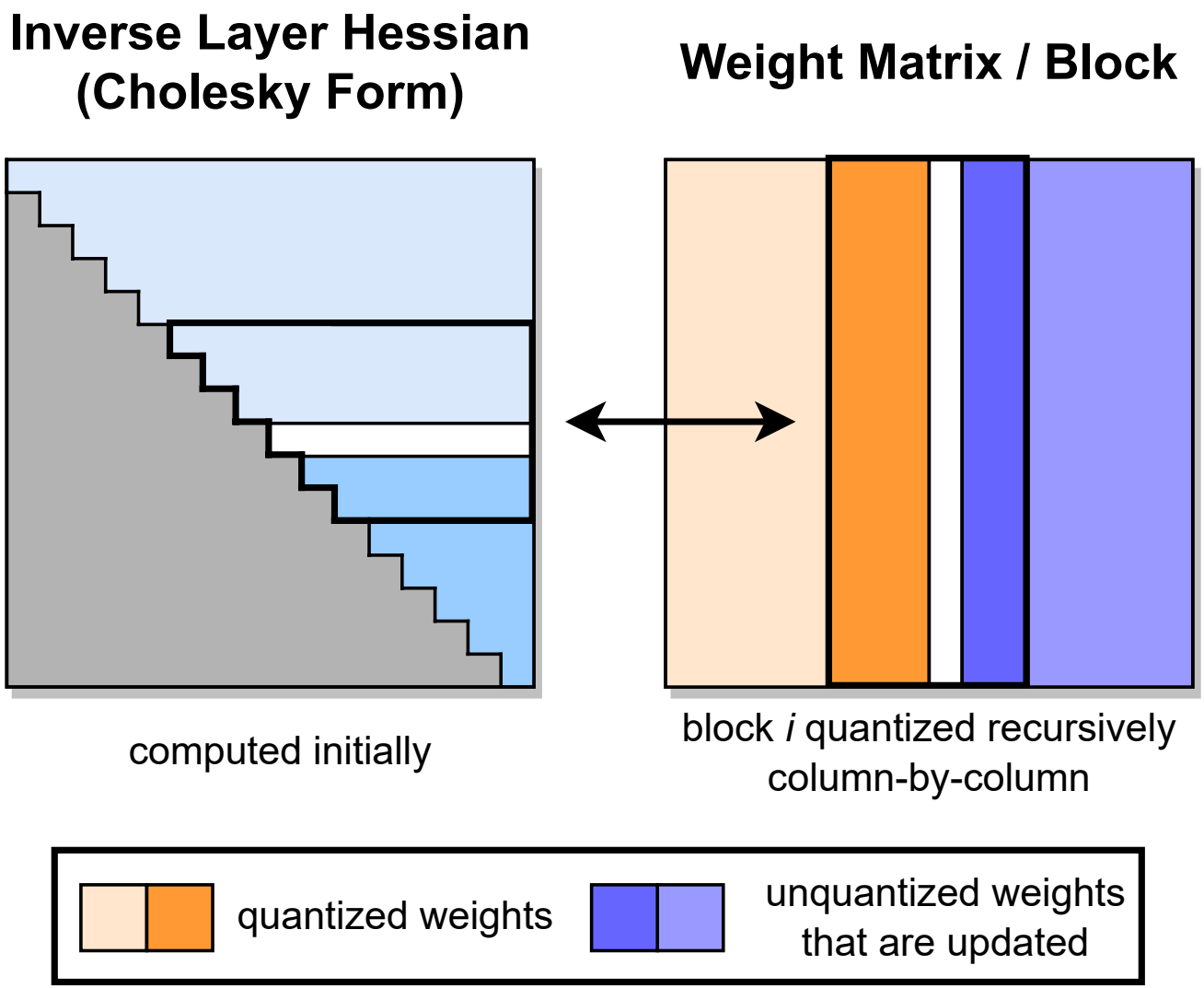
OBQ. Global implementation



GPTQ. Quantization order

Asymptotic improvement – quantize all rows in the same order. $O(d_{row})$ Hessian updates, thus GPTQ time complexity is $O(\max(d_{row}d_{col}^2, d_{col}^3))$. On large models the greedy order of OBQ generally gives only small improvement over an arbitrary order. The speculation is following. The small number of quantized weights with large individual error is balanced out by those weights being quantized towards the end of the process, when only few other unquantized weights that can be adjusted for error compensation remain [4].

GPTQ. Block update



GPTQ. Cholesky reformulation

1. the only information required from $\mathbf{H}_{F_q}^{-1}$, where F_q denotes the set of unquantized weights when quantizing weight q, is the elements in q-th row starting with the diagonal.
2. precompute all of these rows using a more numerically-stable method
3. the row removal for symmetric $\mathbf{H}^{-1}$ corresponds to taking a Cholesky decomposition, except for the minor difference that the latter divides row q by $\sqrt{[\mathbf{H}_{F_q}^{-1}]_{qq}}$.

GPTQ. Algorithm

Algorithm 2 Quantize $\mathbf{W}$ given inverse Hessian $\mathbf{H}^{-1} = (2\mathbf{X}\mathbf{X}^\top + \lambda\mathbf{I})^{-1}$ and blocksize B.

$\mathbf{Q} \leftarrow \mathbf{0}_{d_{\text{row}} \times d_{\text{col}}}$ {quantized output}
$\mathbf{E} \leftarrow \mathbf{0}_{d_{\text{row}} \times B}$ {block quantization errors}
$\mathbf{H}^{-1} \leftarrow \text{Cholesky}(\mathbf{H}^{-1})$ {Hessian inverse information}
for $i = 0, B, 2B, \ldots$ **do**
  for $j = i, \ldots, i + B - 1$ **do**
    $\mathbf{Q}_{:,j} \leftarrow \text{quant}(\mathbf{W}_{:,j})$ {quantize column}
    $\mathbf{E}_{:,j-i} \leftarrow (\mathbf{W}_{:,j} - \mathbf{Q}_{:,j}) \,/\, [\mathbf{H}^{-1}]_{jj}$ {quantization error}
    $\mathbf{W}_{:,j:(i+B)} \leftarrow \mathbf{E}_{:,j-i} \cdot \mathbf{H}^{-1}_{j,j:(i+B)}$ {update weights in block}
  end for
  $\mathbf{W}_{:,(i+B):} \leftarrow \mathbf{E} \cdot \mathbf{H}^{-1}_{i:(i+B),(i+B):}$ {update all remaining weights}
end for

Experiments

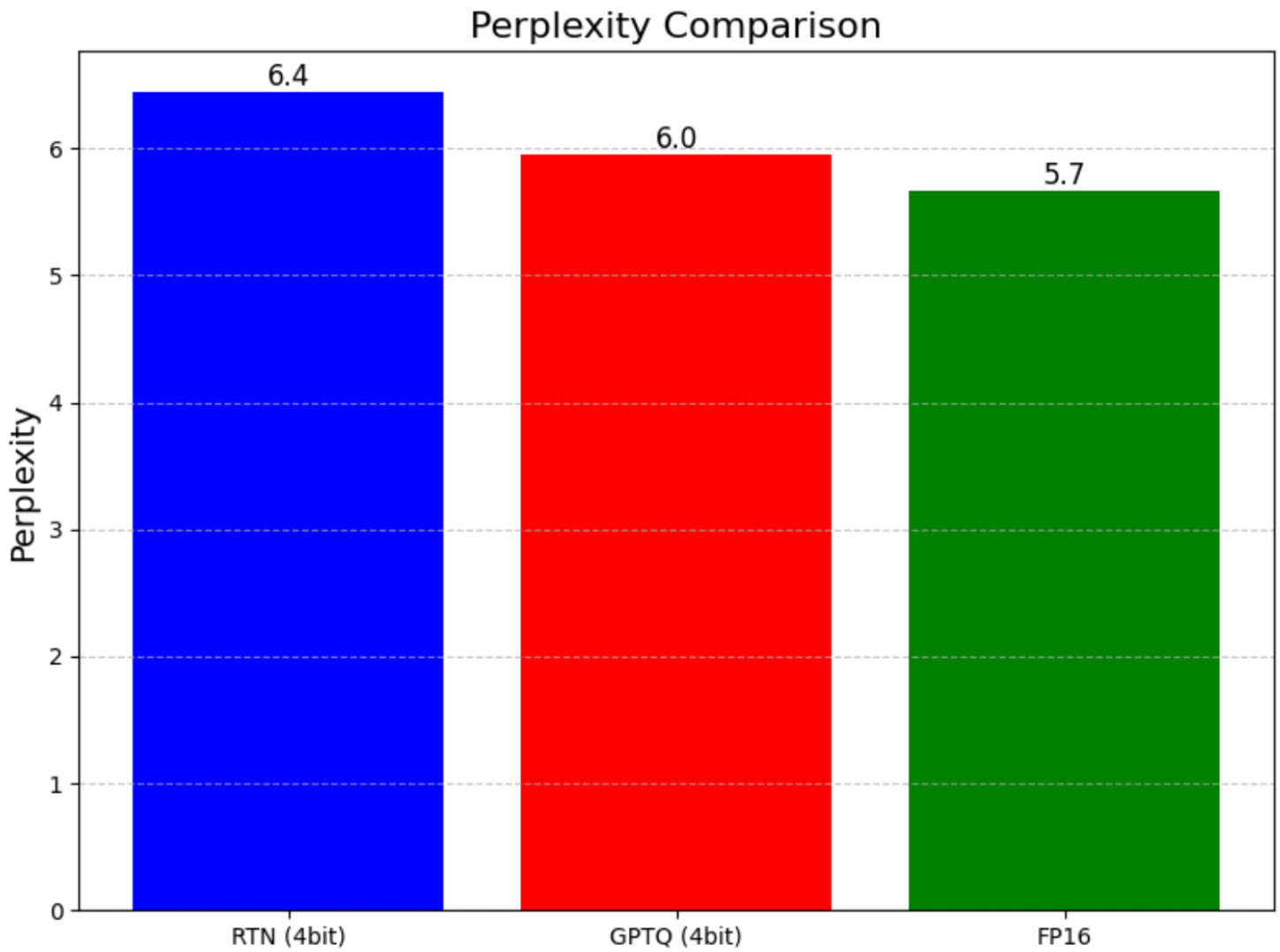


Figure: Perplexity of 4 bit quantizations RTN, GPTQ, and half-precision models.

Literature

[1] Y. LeCun, J. Denker, and S. Solla, "Optimal brain damage," in Advances in Neural Information Processing Systems, D. Touretzky, Ed., vol. 2, Morgan-Kaufmann, 1989. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/1989/file/6c9882bbac1c7093bd25041881277658-Paper.pdf.

[2] B. Hassibi and D. Stork, "Second order derivatives for network pruning,", 1993.

[3] E. Frantar, S. Pal Singh, and D. Alistarh, "Optimal brain compression: A framework for accurate post-training quantization and pruning,", 2021.

[4] S. Ashkboos, T. Hoefler, and D. Alistarh, "Gptq: Accurate post-training quantization for generative pre-trained transformers,", 2022.