Optimal Brain Compression: A Framework for Accurate Post-Training Quantization and Pruning

NeurIPS 2022

Abstract

- A challenging one-shot/post-training setting, in which we are given an accurate trained model, and must compress it without any retraining, based only on a small amount of calibration input data.
- A new compression framework which covers both weight pruning and quantization in a unified setting, is time and space-efficient, and considerably improves upon the practical performance of existing post-training methods
- Our experimental results show that it can improve significantly upon the compression-accuracy trade-offs of existing post-training methods, and that it can enable the accurate compound application of both pruning and quantization in a post-training setting.

Frantar, Elias, and Dan Alistarh. "Optimal Brain Compression: A framework for accurate post-training quantization and pruning." arXiv preprint arXiv:2208.11580 (2022).

Base idea

- Our approach based on an exact and efficient realization of the classical Optimal Brain Surgeon (OBS) framework of [LeCun, Denker, and Solla, 1990] extended to also cover weight quantization at the scale of modern DNNs

Frantar, Elias, and Dan Alistarh. "Optimal Brain Compression: A framework for accurate post-training quantization and pruning." arXiv preprint arXiv:2208.11580 (2022).

Introduction

1. Parameters => More, Model => Larger
2. Limitation 1 => expensive retraining
3. Limitation 2 => Pruning and Quantization independently
4. New setup => MLPerf Close Trip => No Retraining => Post-training Compression Setup => one-shot => limited computational cost
5. Optimal Brain Surgeon => Choose removed weights by leveraging second-order information
6. Limitation 3 => OBS can lead to state-of-the-art compression at DNN scale, by introducing numerical methods which can approximate the second-order information needed by OBS at the massive parameter counts of modern models. However, these approaches do not apply to the post-training setting, as they require gradual pruning, as well as significant retraining, to recover good accuracy

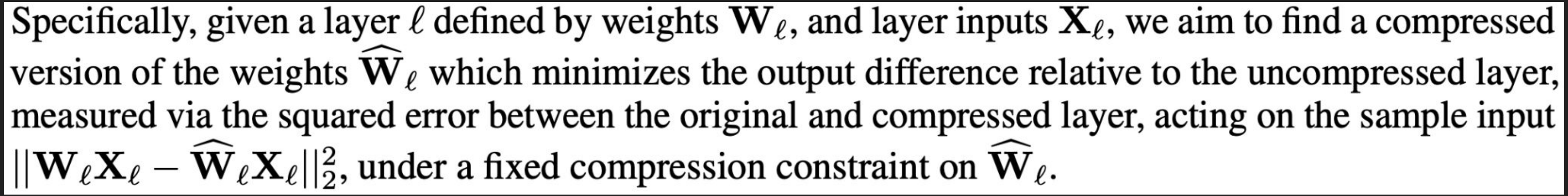
Frantar, Elias, and Dan Alistarh. "Optimal Brain Compression: A framework for accurate post-training quantization and pruning." arXiv preprint arXiv:2208.11580 (2022).

- Post-training Compression => Layer-wise Sub-problems => Compressed weight approximation for each layer => Calibration data
- Layerwise Quantization + Layerwise Pruning
- Limitation 4 => Whether existing approaches for pruning and quantization can be unified in order to cover both types of compression in the post-training setting.

Frantar, Elias, and Dan Alistarh. "Optimal Brain Compression: A framework for accurate post-training quantization and pruning." arXiv preprint arXiv:2208.11580 (2022).

Contribution

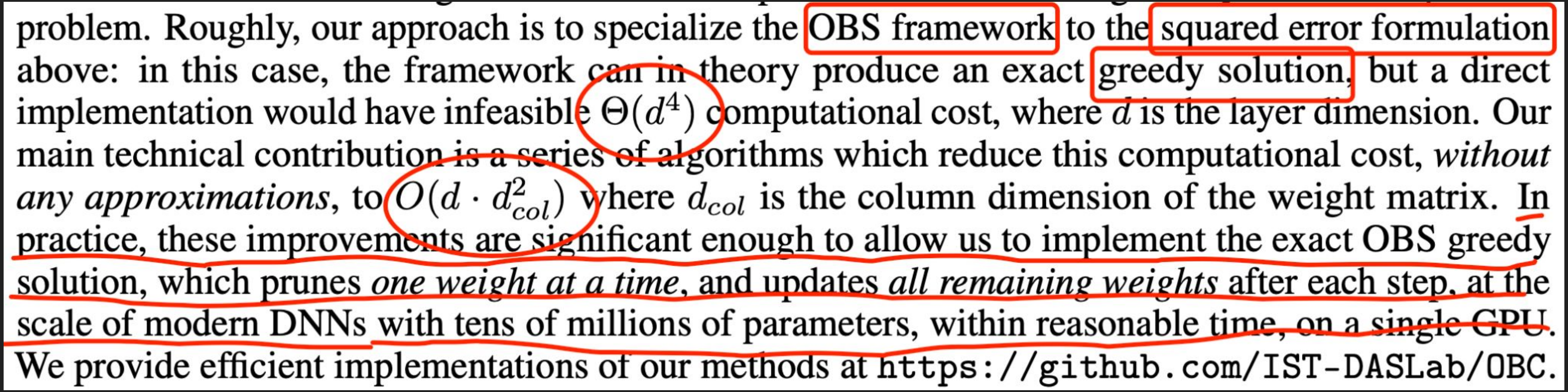
1. In this paper, we provide a mathematical framework for compression via pruning or quantization, which leads to state-of-the-art accuracy-versus-compression trade-offs in the challenging post-training compression setup.

2. 

Specifically, given a layer ℓ defined by weights $\mathbf{W}_\ell$, and layer inputs $\mathbf{X}_\ell$, we aim to find a compressed version of the weights $\widehat{\mathbf{W}}_\ell$ which minimizes the output difference relative to the uncompressed layer, measured via the squared error between the original and compressed layer, acting on the sample input $||\mathbf{W}_\ell \mathbf{X}_\ell - \widehat{\mathbf{W}}_\ell \mathbf{X}_\ell||_2^2$, under a fixed compression constraint on $\widehat{\mathbf{W}}_\ell$.

3. Optimal Brain Quantizer (OBQ), quantized weights iteratively one-ata-time, depending on their impact on the loss increase, after which it applies a closed-form update to the remaining unquantized weights, further reducing the loss

Frantar, Elias, and Dan Alistarh. "Optimal Brain Compression: A framework for accurate post-training quantization and pruning." arXiv preprint arXiv:2208.11580 (2022).



problem. Roughly, our approach is to specialize the $\boxed{\text{OBS framework}}$ to the $\boxed{\text{squared error formulation}}$ above: in this case, the framework can in theory produce an exact $\boxed{\text{greedy solution}}$, but a direct implementation would have infeasible $\Theta(d^4)$ computational cost, where d is the layer dimension. Our main technical contribution is a series of algorithms which reduce this computational cost, *without any approximations*, to $O(d \cdot d_{col}^2)$ where d_{col} is the column dimension of the weight matrix. In practice, these improvements are significant enough to allow us to implement the exact OBS greedy solution, which prunes *one weight at a time*, and updates *all remaining weights* after each step, at the scale of modern DNNs with tens of millions of parameters, within reasonable time, on a single GPU. We provide efficient implementations of our methods at `https://github.com/IST-DASLab/OBC`.

Frantar, Elias, and Dan Alistarh. "Optimal Brain Compression: A framework for accurate post-training quantization and pruning." arXiv preprint arXiv:2208.11580 (2022).

- we show that our pruning and quantization approaches can be compounded, with surprisingly strong results: we obtain a 12× reduction in theoretical operations with a 2% accuracy drop for GPU-supported compound compression [30], 2 and a 4× speedup in actual runtime with only 1% accuracy drop for a CPU-based sparsity-aware runtime [35]. Together, these results suggest for the first time that post-training compression can be competitive with full retraining

Frantar, Elias, and Dan Alistarh. "Optimal Brain Compression: A framework for accurate post-training quantization and pruning." arXiv preprint arXiv:2208.11580 (2022).

Related work

- Optimal Brain Surgeon (OBS).
- Post-Training Quantization.
- Post-Training Sparsification.
- Non-Uniform Compression.

Frantar, Elias, and Dan Alistarh. "Optimal Brain Compression: A framework for accurate post-training quantization and pruning." arXiv preprint arXiv:2208.11580 (2022).

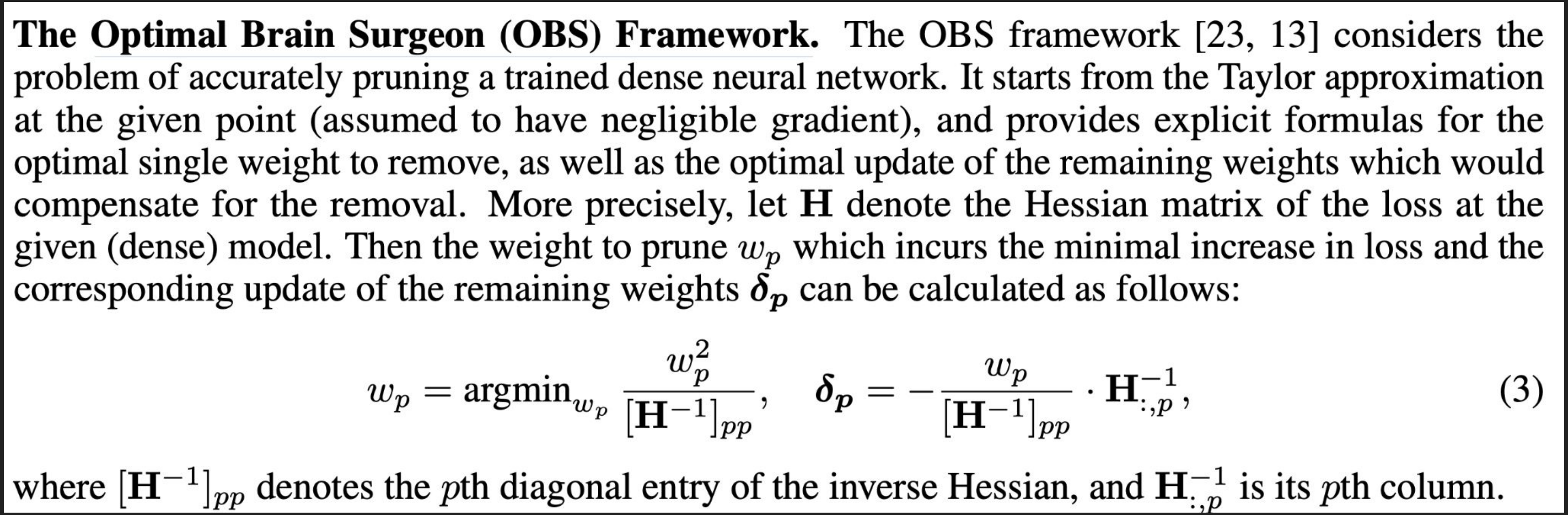
Methodology - Layerwise Compression Problem.

$$\text{argmin}_{\widehat{W}_\ell} \quad \mathbb{E}_{X_\ell}\, \mathcal{L}(f_\ell(X_\ell, W_\ell), f_\ell(X_\ell, \widehat{W}_\ell)) \quad \text{subject to} \quad \mathcal{C}(\widehat{W}_\ell) > C.$$

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

Frantar, Elias, and Dan Alistarh. "Optimal Brain Compression: A framework for accurate post-training quantization and pruning." arXiv preprint arXiv:2208.11580 (2022).

Methodology - Optimal Brain Surgeon (OBS) Framework



The Optimal Brain Surgeon (OBS) Framework. The OBS framework [23, 13] considers the problem of accurately pruning a trained dense neural network. It starts from the Taylor approximation at the given point (assumed to have negligible gradient), and provides explicit formulas for the optimal single weight to remove, as well as the optimal update of the remaining weights which would compensate for the removal. More precisely, let $\mathbf{H}$ denote the Hessian matrix of the loss at the given (dense) model. Then the weight to prune w_p which incurs the minimal increase in loss and the corresponding update of the remaining weights $\boldsymbol{\delta_p}$ can be calculated as follows:

$$w_p = \text{argmin}_{w_p} \frac{w_p^2}{[\mathbf{H}^{-1}]_{pp}}, \quad \boldsymbol{\delta_p} = -\frac{w_p}{[\mathbf{H}^{-1}]_{pp}} \cdot \mathbf{H}^{-1}_{:,p}, \tag{3}$$

where $[\mathbf{H}^{-1}]_{pp}$ denotes the pth diagonal entry of the inverse Hessian, and $\mathbf{H}^{-1}_{:,p}$ is its pth column.

Frantar, Elias, and Dan Alistarh. "Optimal Brain Compression: A framework for accurate post-training quantization and pruning." arXiv preprint arXiv:2208.11580 (2022).

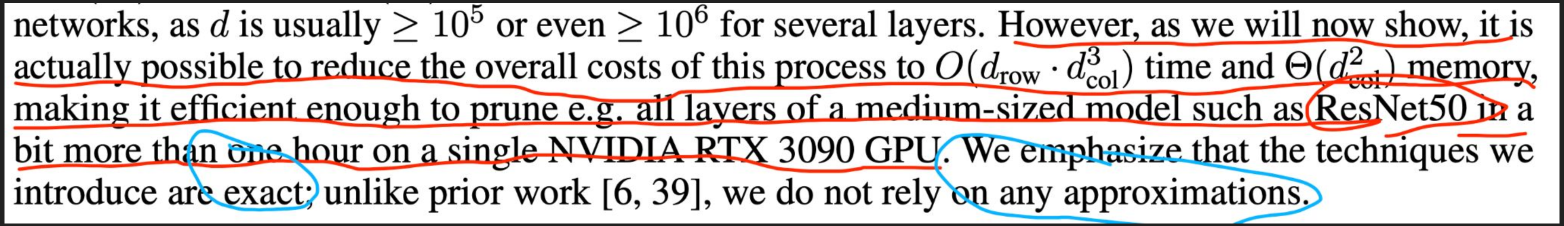
Methodology - OBS for Layer-Wise Pruning

- Instantiate this framework for the layer-wise pruning problem, defined in previous slides.
- Iterating the OBS framework to remove one weight at a time would yield an exact greedy solution for the layer-wise pruning problem, as it takes the (locally) optimal decision at each step.
- While this greedy approach does not guarantee convergence to a global optimum, such approaches can be very effective for dealing with problem instances that are too large to be handled by exact methods.

Frantar, Elias, and Dan Alistarh. "Optimal Brain Compression: A framework for accurate post-training quantization and pruning." arXiv preprint arXiv:2208.11580 (2022).

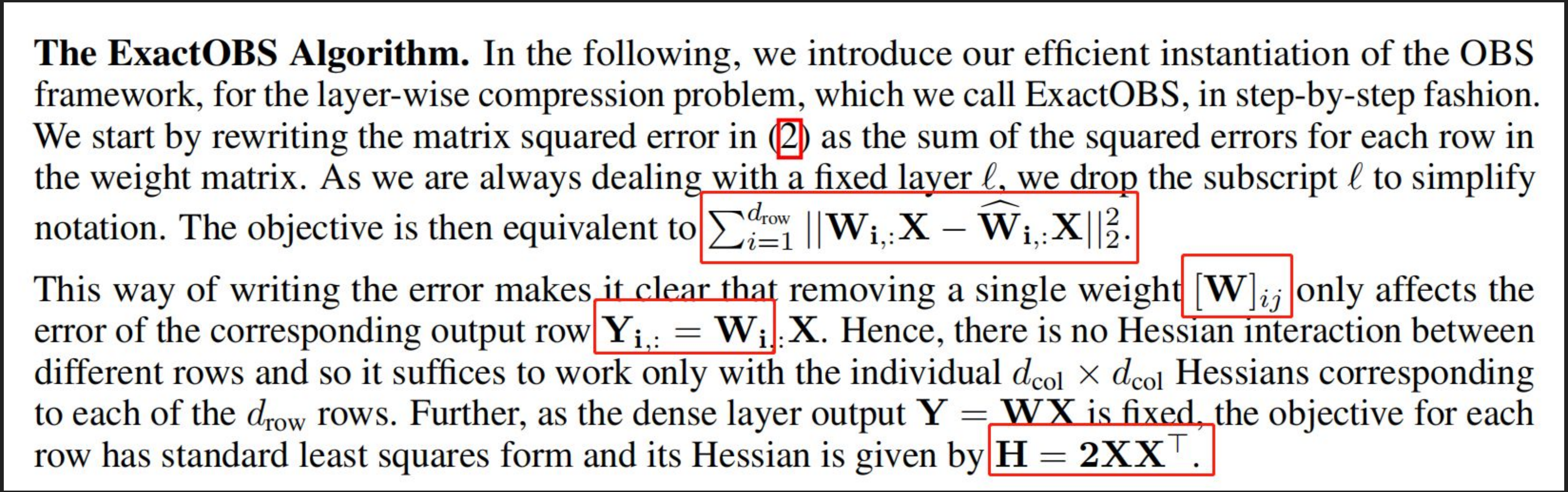
Methodology - An Optimal Greedy Solver for Sparsity

- The Hessian H is a d×d matrix where d = d_row · d_col, which is already expensive to store and compute with
- Additionally, this matrix needs to be updated and inverted at each of the O(d) steps with a computational complexity of Θ(d^3)
- Clearly, an O(d^4) total runtime is too inefficient for pruning most layers of modern neural networks, as d is usually ≥ 10^5 or even ≥ 10^6 for several layers



networks, as d is usually $\geq 10^5$ or even $\geq 10^6$ for several layers. However, as we will now show, it is actually possible to reduce the overall costs of this process to $O(d_{\text{row}} \cdot d_{\text{col}}^3)$ time and $\Theta(d_{\text{col}}^2)$ memory, making it efficient enough to prune e.g. all layers of a medium-sized model such as ResNet50 in a bit more than one hour on a single NVIDIA RTX 3090 GPU. We emphasize that the techniques we introduce are exact; unlike prior work [6, 39], we do not rely on any approximations.

Methodology - An Optimal Greedy Solver for Sparsity

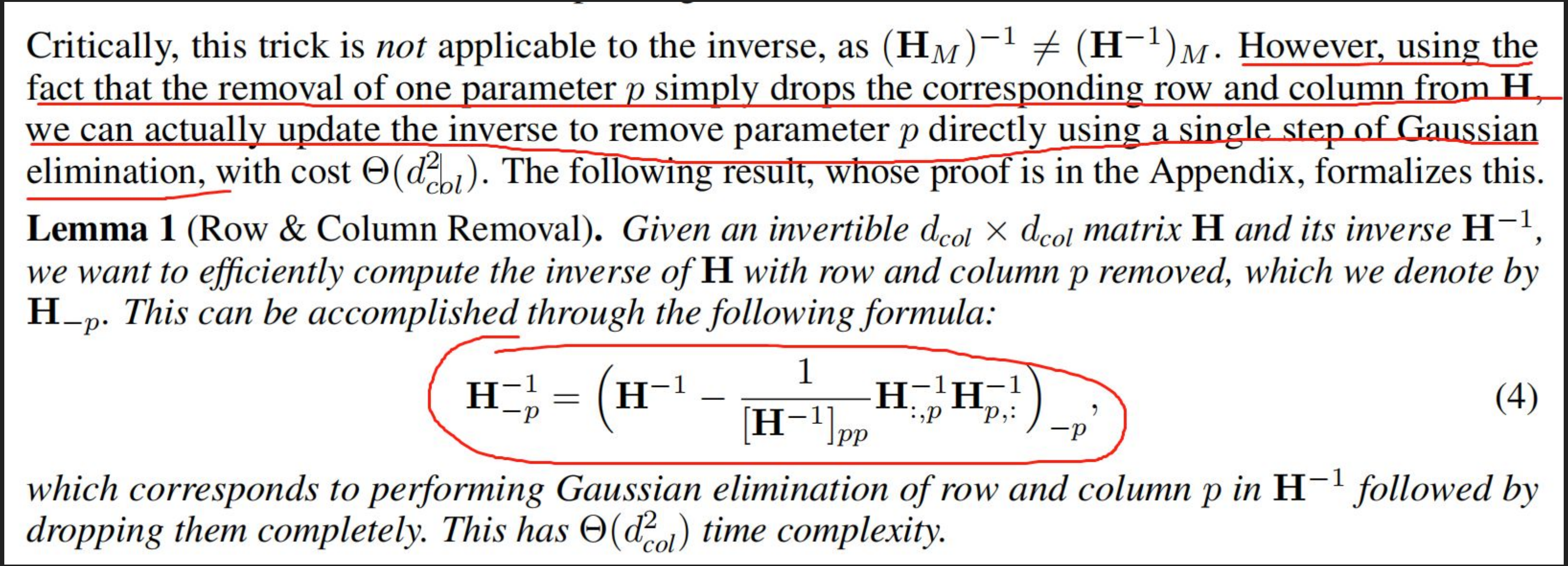


The ExactOBS Algorithm. In the following, we introduce our efficient instantiation of the OBS framework, for the layer-wise compression problem, which we call ExactOBS, in step-by-step fashion. We start by rewriting the matrix squared error in ($\boxed{2}$) as the sum of the squared errors for each row in the weight matrix. As we are always dealing with a fixed layer ℓ, we drop the subscript ℓ to simplify notation. The objective is then equivalent to $\boxed{\sum_{i=1}^{d_{\text{row}}} ||\mathbf{W}_{\mathbf{i},:}\mathbf{X} - \widehat{\mathbf{W}}_{\mathbf{i},:}\mathbf{X}||_2^2.}$

This way of writing the error makes it clear that removing a single weight $\boxed{[\mathbf{W}]_{ij}}$ only affects the error of the corresponding output row $\boxed{\mathbf{Y}_{\mathbf{i},:} = \mathbf{W}_{\mathbf{i},:}}\mathbf{X}$. Hence, there is no Hessian interaction between different rows and so it suffices to work only with the individual $d_{\text{col}} \times d_{\text{col}}$ Hessians corresponding to each of the d_{row} rows. Further, as the dense layer output $\mathbf{Y} = \mathbf{W}\mathbf{X}$ is fixed, the objective for each row has standard least squares form and its Hessian is given by $\boxed{\mathbf{H} = \mathbf{2XX}^\top.}$

Frantar, Elias, and Dan Alistarh. "Optimal Brain Compression: A framework for accurate post-training quantization and pruning." arXiv preprint arXiv:2208.11580 (2022).

Methodology - An Optimal Greedy Solver for Sparsity



Critically, this trick is *not* applicable to the inverse, as $(\mathbf{H}_M)^{-1} \neq (\mathbf{H}^{-1})_M$. However, using the fact that the removal of one parameter p simply drops the corresponding row and column from $\mathbf{H}$, we can actually update the inverse to remove parameter p directly using a single step of Gaussian elimination, with cost $\Theta(d_{col}^2)$. The following result, whose proof is in the Appendix, formalizes this.

Lemma 1 (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:*

$$\boxed{\mathbf{H}_{-p}^{-1} = \left(\mathbf{H}^{-1} - \frac{1}{[\mathbf{H}^{-1}]_{pp}} \mathbf{H}_{:,p}^{-1} \mathbf{H}_{p,:}^{-1}\right)_{-p},} \quad (4)$$

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.

Frantar, Elias, and Dan Alistarh. "Optimal Brain Compression: A framework for accurate post-training quantization and pruning." arXiv preprint arXiv:2208.11580 (2022).

Methodology - An Optimal Greedy Solver for Sparsity

Algorithm 1 Prune $k \leq d_{\text{col}}$ 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
    $p \leftarrow \text{argmin}_{p \in M} \frac{1}{[\mathbf{H}^{-1}]_{pp}} \cdot w_p^2$
    $\mathbf{w} \leftarrow \mathbf{w} - \mathbf{H}^{-1}_{:,p} \frac{1}{[\mathbf{H}^{-1}]_{pp}} \cdot w_p$
    $\mathbf{H}^{-1} \leftarrow \mathbf{H}^{-1} - \frac{1}{[\mathbf{H}^{-1}]_{pp}} \mathbf{H}^{-1}_{:,p} \mathbf{H}^{-1}_{p,:}$
    $M \leftarrow M - \{p\}$
end for
```

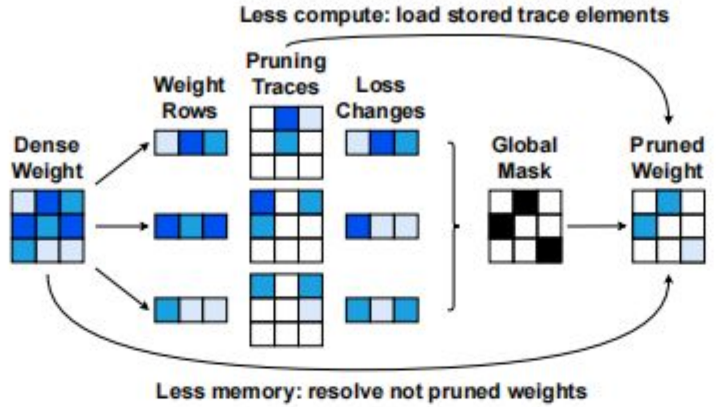

Less compute: load stored trace elements
Weight Rows
Pruning Traces
Loss Changes
Dense Weight
Global Mask
Pruned Weight
Less memory: resolve not pruned weights

Figure 1: Efficient global OBS using the row-wise results.

Frantar, Elias, and Dan Alistarh. "Optimal Brain Compression: A framework for accurate post-training quantization and pruning." arXiv preprint arXiv:2208.11580 (2022).

Thank You