AWP: Activation-Aware Weight Pruning and Quantization with Projected Gradient Descent

Jing Liu, Toshiaki Koike-Akino, Ye Wang, Hassan Mansour, Mathew Brand

Mitsubishi Electric Research Laboratories (MERL), Cambridge, MA 02139, USA

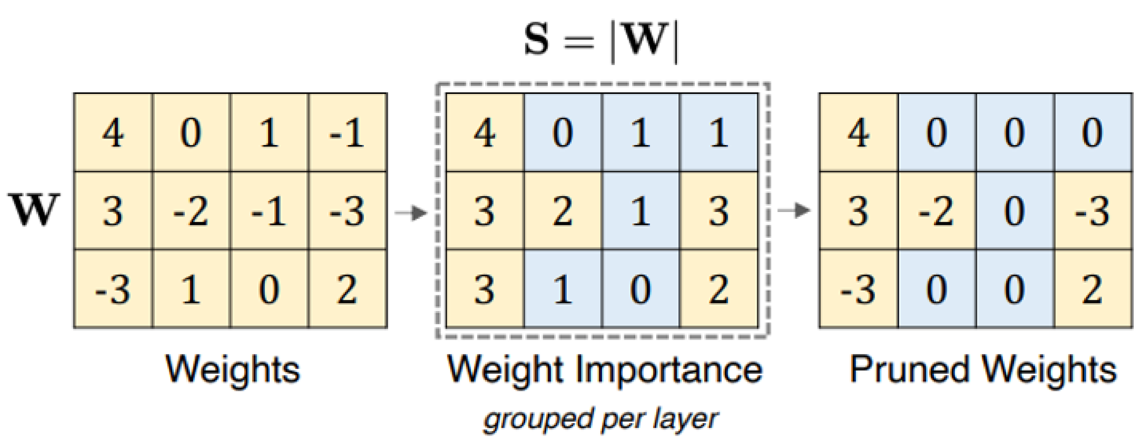
Highlights

- We pose the LLM layer-wise pruning problem as a sparse approximation problem.
- We propose Activation-aware PGD for pruning and/or quantization without requiring computationally-intensive operations such as second-order Hessian inverses.
- The proposed method outperforms state-of-the-art LLM compression methods on several benchmarks.
- We provide theoretical guarantees for the proposed method.

Background and Motivation

$$\min_{\mathbf{W}_{\text{sparse}} \in \mathcal{C}_{\text{sparse}}} \left[\mathcal{L}(\mathbf{W}_{\text{sparse}}) := \|\mathbf{W} - \mathbf{W}_{\text{sparse}}\|_F^2 \right], \quad (1)$$

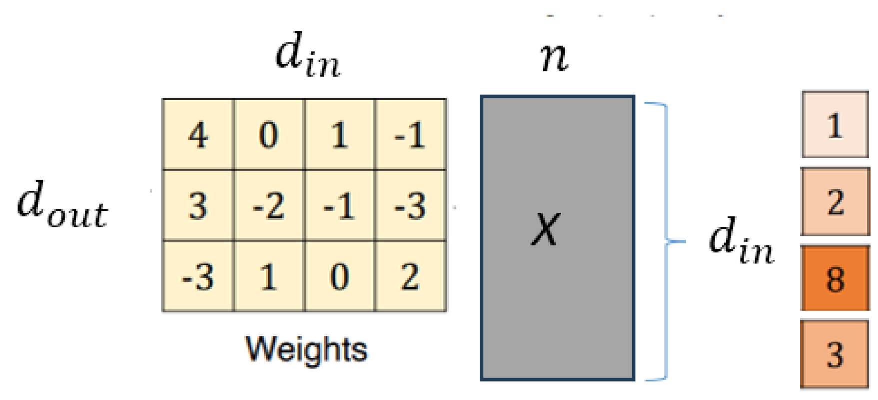
Magnitude Pruning

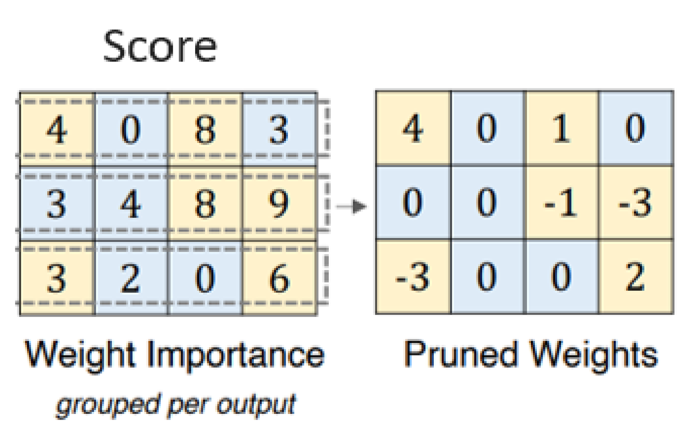


Activation-aware Pruning

$$\mathcal{L}'(\mathbf{W}_{\text{sparse}}) := \|\mathbf{W}\mathbf{X} - \mathbf{W}_{\text{sparse}}\mathbf{X}\|_F^2 \quad (2)$$

Wanda (ICLR'24):





Method

$$\mathcal{L}'(\mathbf{W}_{\text{sparse}}) := \|\mathbf{W}\mathbf{X} - \mathbf{W}_{\text{sparse}}\mathbf{X}\|_F^2 \quad (3)$$
$$= \|\mathbf{W}\mathbf{C}^{\frac{1}{2}} - \mathbf{W}_{\text{sparse}}\mathbf{C}^{\frac{1}{2}}\|_F^2, \quad (4)$$

where $\mathbf{C} = \mathbf{X}\mathbf{X}^\top \in \mathbb{R}^{d_{\text{in}} \times d_{\text{in}}}$ is the auto-correlation of input activation, and $\mathbf{C}^{\frac{1}{2}}$ is the matrix square root of $\mathbf{C}$.
We further decompose (4) as

$$\mathcal{L}'(\mathbf{W}_{\text{sparse}}) = \sum_{i=1}^{d_{\text{out}}} \|\mathbf{W}[i,:]\mathbf{C}^{\frac{1}{2}} - \mathbf{W}_{\text{sparse}}[i,:]\mathbf{C}^{\frac{1}{2}}\|_2^2. \quad (5)$$

when optimizing under the constraint

$$\mathcal{C}_{\text{row}} := \left\{\mathbf{\Theta} : \forall i \in \{1,\ldots,d_{\text{out}}\},\ \|\mathbf{\Theta}[i,:]\|_0 \le k\right\}, \quad (6)$$

Each term of (5) becomes exactly a well-studied sparse approximation problem:

$$\min_{\boldsymbol{\theta}} \left[f(\boldsymbol{\theta}) := \|\mathbf{y} - \mathbf{A}\boldsymbol{\theta}\|_2^2 \right], \quad (7)$$
$$\text{s.t. } \|\boldsymbol{\theta}\|_0 \le k := (1-p)\cdot d_{\text{in}},$$

where $\mathbf{y} = (\mathbf{W}[i,:]\mathbf{C}^{\frac{1}{2}})^\top$, $\mathbf{A} = [\mathbf{C}^{\frac{1}{2}}]^\top = \mathbf{C}^{\frac{1}{2}}$, and $\boldsymbol{\theta}$ is the corresponding $\mathbf{W}_{\text{sparse}}[i,:]^\top$ with k nonzeros.

Inspired by Iterative Hard Thresholding that iterates between gradient descent and hard thresholding.

Algorithm 1: Activation-Aware Projected Gradient Descent

Input: original weight $\mathbf{W} \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$, input activation covariance $\mathbf{C} = \frac{1}{n}\mathbf{X}\mathbf{X}^\top$, constraints $\mathcal{C}$, step size η
Initialize: $\mathbf{\Theta}^{(0)} \in \mathcal{C}$
Repeat
  $\mathbf{Z}^{(t)} = \mathbf{\Theta}^{(t)} + \eta(\mathbf{W} - \mathbf{\Theta}^{(t)})\mathbf{C}$;
  $\mathbf{\Theta}^{(t+1)} = \text{Proj}_{\mathcal{C}}(\mathbf{Z}^{(t)})$;
until a stopping criterion is met
Output compressed weight $\mathbf{\Theta}$

Theoretical guarantees established for the pruning case (see appendix).

Experiments

- **Pruning**

Table: Perplexity on WikiText2 of pruned Llama-2-7B model by different methods under different pruning ratios.

	50%	60%	70%	80%	90%
Magnitude	14.89	4e3	-	NAN	-
SparseGPT	6.51	9.58	-	1e2	-
Wanda	6.48	10.09	70.04	4e3	1e4
AWP	**6.42**	**9.44**	**22.10**	**83.28**	**8e2**

Table: Perplexity on WikiText2 of pruned Llama-2-13B model by different methods under different pruning ratios.

	50%	60%	70%	80%	90%
Magnitude	6.37	11.23	-	5e4	-
SparseGPT	5.63	7.80	-	1e2	-
Wanda	5.59	7.97	43.06	1e3	2e4
AWP	**5.54**	**7.49**	**16.57**	**75.68**	**1e3**

- **Quantization**

Table: Perplexity on WikiText2 of quantized Llama-3.1-8B model by different methods.

	INT4	INT3	INT2
GPTQ	9.95	12.54	**2e3**
AWQ	6.64	8.14	3e4
AWP	**6.55**	**8.06**	1e6

- **Quantization & Pruning**

Table: Perplexity on WikiText2 of pruned and INT4 quantized Llama-3.1-8B model by different methods.

Pruning Ratio:	25%	50%	75%
AWQ+Wanda	6.93	9.71	3e2
Wanda+AWQ	**6.81**	9.46	2e2
AWP	**6.81**	**9.32**	**1e2**

Table: Perplexity on WikiText2 of pruned and INT4 quantized Llama-3.2-1B model by different methods.

Pruning Ratio:	25%	50%	75%
AWQ+Wanda	11.63	23.95	2e3
Wanda+AWQ	11.30	21.90	1e3
AWP	**11.20**	**18.41**	**3e2**

Our Related Work: Activation-aware low-rank compression (LatentLLM, CVPR-W'25)
Naïve SVD of $\mathbf{W}$ would minimize $\|\mathbf{W} - \mathbf{B}\mathbf{A}\|_F^2$.
Activation-aware SVD aims to minimize:

$$\|\mathbf{W}\mathbf{X} - \mathbf{B}\mathbf{A}\mathbf{X}\|_F^2 = \|\mathbf{W}\mathbf{C}^{1/2} - \underbrace{\mathbf{B}\mathbf{A}\mathbf{C}^{1/2}}_{\text{rank} \le r}\|_F^2.$$

Our global optimal solution:
The optimal rank r approximation of $\mathbf{W}\mathbf{C}^{1/2}$ can be obtained by SVD of $\mathbf{W}\mathbf{C}^{1/2} = \mathbf{U}\mathbf{S}\mathbf{V}^T$ and keep its top-r components $\mathbf{U}_r\mathbf{S}_r\mathbf{V}_r^T$.
We can set $\mathbf{B}\mathbf{A}\mathbf{C}^{1/2} = \mathbf{U}_r\mathbf{S}_r\mathbf{V}_r^T$ and obtain $\mathbf{B}\mathbf{A} = \mathbf{U}_r\mathbf{S}_r\mathbf{V}_r^T\mathbf{C}^{-1/2}$
So setting $\mathbf{B} = \mathbf{U}_r$ and $\mathbf{A} = \mathbf{S}_r\mathbf{V}_r^T\mathbf{C}^{-1/2}$ would be a global optimal solution.