Symmetric Pruning for Large Language Models

Kai Yi   Peter Richtarik

King Abdullah University of Science and Technology (KAUST)

Introduction

Background. Large Language Models (LLMs) are growing rapidly in size and capabilities, posing significant computational and memory challenges. **Post-training pruning** (PTP) has emerged as a key method for reducing the footprint of pretrained weights.

Popular PTP methods

- Magnitude-based pruning: elements of each layer's weights with smaller absolute values are set to zero.
- Wanda (Sun et al., 2023): scales the weights by the activations of each layer, demonstrating promising performance on standard benchmarks.
- RIA (Zhang et al., 2024b): further improved the approach by evaluating the relative importance of each weight across its corresponding row and column before pruning.

While their empirical results are encouraging, the underlying mechanisms remain poorly understood. This leads us to our first question:

Q1: Can we provide theoretical support for PTP methods and derive more efficient algorithms with minimal adaptations to the existing framework?

To better understand popular PTP methods, we propose Symmetric Weight And Activation (SymWanda)—a novel formulation that leverages both input activations and layer outputs. This symmetric approach offers theoretical insights into methods like Wanda and RIA.

Training-Free Fine-Tuning

While intrinsic PTP methods achieve strong perplexity and zero-shot accuracy, they struggle at high sparsity due to **reconstruction errors** between pruned and original weights. Minimizing this error is crucial for efficient PTP.

Q2: Can we fine-tune pruned LLMs without further training and outperforms state-of-the-art methods with minimal effort?

Dynamic Sparse Training (DST) efficiently updates a subset of network parameters while adapting the sparse topology during training. Though promising for fine-tuning LLMs, DST depends on backpropagation and frequent weight updates, limiting its efficiency for large-scale models.

We explore **training-free fine-tuning** by leveraging the **pruning-and-growing** mechanism in DST, which adapts sparse masks based on weight properties without backpropagation. DSnoT (Zhang et al., 2023) introduced a simple approach using weight values and statistics but struggles with non-uniform weight distributions. To overcome this, we:

- Incorporate relative weight importance in mask updates.
- Introduce a regularization term to better optimize reconstruction error.

Contribution

- We propose SymWanda, a novel formulation that reduces pruning impact on input activations and output influences, offering theoretical insights into Wanda and RIA.
- Based on SymWanda, we develop new pruning strategies, validated through extensive experiments. An efficient stochastic approach for relative importance manipulation achieves superior performance with reduced sampling cost.
- We introduce R^2-DSnoT, a training-free fine-tuning method that leverages relative weight importance and a regularized decision boundary in a pruning-and-growing framework, significantly outperforming strong baselines.

Symmetric Wanda: New Formulations

Table 1. Comparison of LLM post-training pruning algorithms.

Algorithm	W?	Act.?	$\mathbf{X}$	$\mathbf{Y}$	$\mathbf{S}_{jk}$ [(a)]
General Sym.	✓	✓	$\mathbf{X}$	$\mathbf{Y}$	$\lvert\mathbf{W}_{jk}\rvert\,(\lVert\mathbf{X}_{:j}\rVert_2+\lVert\mathbf{Y}_{k:}\rVert_2)$
Marginal	✓	✗	$\mathbf{I}$	$\mathbf{0}$	$\lvert\mathbf{W}_{jk}\rvert$
Wanda	✓	✓	$\mathbf{X}$	$\mathbf{0}$	$\lvert\mathbf{W}_{jk}\rvert\lVert\mathbf{X}_{:j}\rVert_2$
OWanda	✓	✓	$\mathbf{0}$	$\mathbf{Y}$	$\lvert\mathbf{W}_{jk}\rvert\lVert\mathbf{Y}_{k:}\rVert_2$
Symmetric	✓	✓	$\mathbf{W}^T$	$\mathbf{W}^T$	$\lvert\mathbf{W}_{jk}\rvert\sqrt{\lVert\mathbf{W}_{j:}\rVert_2^2+\lVert\mathbf{W}_{:k}\rVert_2^2}$
RI (v1)	✓	✗	$t_j(1;,\cdots;,1),\ t_j=(\sqrt{b}\lVert\mathbf{W}_{j:}\rVert_1)^{-1}$ [(a)]	$s_k(1,\cdots,1),\ s_k=(\sqrt{c}\lVert\mathbf{W}_{:k}\rVert_1)^{-1}$	$\lVert\mathbf{W}_{j:}\rVert_1^{-1}+\lVert\mathbf{W}_{:k}\rVert_1^{-1}$
RI (v2)	✓	✗	$(\lVert\mathbf{W}_{1:}\rVert_1^{-1},\ldots,\lVert\mathbf{W}_{b:}\rVert_1^{-1})$	$(\lVert\mathbf{W}_{:1}\rVert_1^{-1},\ldots,\lVert\mathbf{W}_{:c}\rVert_1^{-1})$	$\lVert\mathbf{W}_{j:}\rVert_1^{-1}+\lVert\mathbf{W}_{:k}\rVert_1^{-1}$
RIA	✓	✓	$\delta_{u=j}\delta_{v=p}\lVert\mathbf{C}_{:j}\rVert_2^{\alpha}\lVert\mathbf{W}_{j:}\rVert_1^{-1}$ [(c)]	$\delta_{u=s}\delta_{v=k}\lVert\mathbf{C}_{:j}\rVert_2^{\alpha}\lVert\mathbf{W}_{:k}\rVert_1^{-1}$	$\left(\lVert\mathbf{W}_{j:}\rVert_1^{-1}+\lVert\mathbf{W}_{:k}\rVert_1^{-1}\right)\lVert\mathbf{X}_{:j}\rVert_2^{\alpha}$
General (diag.)	✓	✓	$\mathbf{A}\mathbf{D}_{\mathbf{X}}$ [(d)]	$\mathbf{D}_{\mathbf{Y}}\mathbf{B}$	$\lVert\mathbf{A}_{:j}\rVert_2\lVert\mathbf{W}_{j:}\rVert_1^{-1}+\lVert\mathbf{B}_{k:}\rVert_2\lVert\mathbf{W}_{:k}\rVert_1^{-1}$
ℓ_p-norm (v1)	✓	✗ [(e)]	$\lVert\mathbf{W}_{j:}\rVert_p^{-1}\cdot\lVert\mathbf{W}_{j:}\rVert_2^{-1}\cdot\mathbf{W}_{j:}^{\top}$	$\lVert\mathbf{W}_{:k}\rVert_p^{-1}\cdot\lVert\mathbf{W}_{:k}\rVert_2^{-1}\cdot\mathbf{W}_{:k}^{\top}$	$\lvert\mathbf{W}_{jk}\rvert(\lVert\mathbf{W}_{j:}\rVert_p^{-1}+\lVert\mathbf{W}_{:k}\rVert_p^{-1})$
ℓ_p-norm (v2)	✓	✗	$\lVert\mathbf{W}_{j:}\rVert_p^{-1}\cdot\mathbf{u}$	$\lVert\mathbf{W}_{:k}\rVert_p^{-1}\cdot\mathbf{v}$	$\lvert\mathbf{W}_{jk}\rvert(\lVert\mathbf{W}_{j:}\rVert_p^{-1}+\lVert\mathbf{W}_{:k}\rVert_p^{-1})$
StochRIA	✓	✗	$\mathbf{1}_{\{i\in S_j\}}\left(\lVert\mathbf{W}_{j:S_j}\rVert_1\sqrt{\tau}\right)^{-1}$	$\mathbf{1}_{\{i\in S_k\}}\left(\lVert\mathbf{W}_{S_k:k}\rVert_1\sqrt{\tau}\right)^{-1}$	$\lvert\mathbf{W}_{jk}\rvert(\lVert\mathbf{W}_{j:S_j}\rVert_1^{-1}+\lVert\mathbf{W}_{S_k:k}\rVert_1^{-1})$

[(a)] Without loss of generality, we consider the elimination of a single weight, $\mathbf{W}_{jk}$. The detailed explanation can be found in Lemma 3.1 and Section 3.2.
[(b)] For simplicity, instead of displaying the entire matrices $\mathbf{X}$ and $\mathbf{Y}$, we present the columns $\mathbf{X}_{:j}$ and the rows $\mathbf{Y}_{k:}$.
This design is employed in the algorithms RI, RIA, ℓ_p-norm, and StochRIA.
[(c)] The Kronecker delta, denoted by δ_{ij}, is a function of two indices i and j that equals 1 if $i=j$ and 0 otherwise.
[(d)] $\mathbf{D}_{\mathbf{X}}$ and $\mathbf{D}_{\mathbf{Y}}$ are the diagonal matrices associated with $\mathbf{W}$, as defined in Section 3.4.
[(e)] By default, for ℓ_p-norm and StochRIA, we do not consider the input activation. However, the design is similar to the transition from RI to RIA, as described in Section 3.3.

Consider a target sparsity ratio $\varepsilon\in[0,1)$, a set of calibration inputs $\mathbf{X}\in\mathbb{R}^{a\times b}$, and pre-trained weights $\mathbf{W}\in\mathbb{R}^{b\times c}$. The objective is to identify an optimal pruned weight matrix $\widetilde{\mathbf{W}}\in\mathbb{R}^{b\times c}$ that minimizes:

$$f(\widetilde{\mathbf{W}}) := \lVert\mathbf{X}(\widetilde{\mathbf{W}}-\mathbf{W})\rVert_F^2, \qquad \text{(InpRecon)}$$

where the optimization challenge is: $\text{minimize } f(\widetilde{\mathbf{W}})$ s.t. $\text{Mem}(\widetilde{\mathbf{W}})\le(1-\varepsilon)\text{Mem}(\mathbf{W})$.

Apart from the previous defined input calibration $\mathbf{X}$, we particularly introduce the output calibration $\mathbf{Y}\in\mathbb{R}^{c\times d}$. Considering both the input and output dependencies, we express the objective as:

$$g(\widetilde{\mathbf{W}}) := \lVert\mathbf{X}(\widetilde{\mathbf{W}}-\mathbf{W})\rVert_F+\lVert(\widetilde{\mathbf{W}}-\mathbf{W})\mathbf{Y}\rVert_F, \qquad \text{(Sym)}$$

and propose to solve: $\text{minimize }\ g(\widetilde{\mathbf{W}})$, s.t. $\text{Mem}(\widetilde{\mathbf{W}})\le(1-\varepsilon)\text{Mem}(\mathbf{W})$.

Key Lemma: Assume we aim to eliminate a single weight $\mathbf{W}_{jk}$, setting $\widetilde{\mathbf{W}}_{jk}=0$ and keeping all other weights unchanged. The simplified expression for $g(\widetilde{\mathbf{W}})$ becomes:

$$g(\widetilde{\mathbf{W}})=\lvert\mathbf{W}_{jk}\rvert\,(\lVert\mathbf{X}_{:j}\rVert_2+\lVert\mathbf{Y}_{k:}\rVert_2):=\mathbf{S}_{jk}.$$

Relative and Regularized Dynamic Sparse no Training (R^2-DSnoT)

Define $\mathbf{D}_{q,r}:=\lVert\widetilde{\mathbf{W}}_{q,:}\rVert_1^{-1}+\lVert\widetilde{\mathbf{W}}_{:,r}\rVert_1^{-1}$. The updated rule for identifying the growing index i is formalized as:

$$i=\arg\max_{r}\left\{\text{sign}(\mathbb{E}[\epsilon_q])\cdot\mathbf{D}_{q,r}\cdot\frac{\mathbb{E}[\mathbf{X}_q]}{\text{Var}(\mathbf{X}_q)}+\gamma_1\lVert\widetilde{\mathbf{W}}_q\rVert_p\right\}, \qquad (1)$$

where γ_1 is the growing regularization parameter, striking a balance between fidelity and the ℓ_p regularizer. Similarly, the pruning index j is now defined as:

$$j=\arg\min_{r:\Delta(q,r)<0}\left\{\lvert\widetilde{\mathbf{W}}_{q,r}\rvert\cdot\mathbf{D}_{q,r}\cdot\lVert\mathbf{X}_q\rVert_2^{\alpha}+\gamma_2\lVert\widetilde{\mathbf{W}}_q\rVert_p\right\}, \qquad (2)$$

where $\Delta(q,r):=\text{sign}\big(\mathbb{E}[\epsilon_q]\big)\left(\widetilde{\mathbf{W}}q,r\cdot\mathbf{D}q,r\cdot\mathbb{E}[\mathbf{X}_q]\right)$, and γ_2 denotes the pruning regularization parameter.

Experiments

Sparsity	Method	Sampling	LlaMA2-7b	LlaMA2-13b	LlaMA3-8b	OPT-1.3b
-	Dense	-	5.47	4.88	6.14	14.62
50%	Magnitude	-	16.03	6.83	205.44	1712.39
	Wanda	-	7.79	6.28	10.81	22.19
	RIA	Full	6.88	5.95	9.44	18.94
	stochRIA	10%	$6.91^{\pm 0.0032}_{-0.03}$	$5.95^{\pm 0.0033}_{+0}$	$9.46^{\pm 0.025}_{-0.02}$	$18.78^{\pm 0.050}_{+0.16}$
2:4	RIA	Full	11.31	8.40	22.89	27.43
	stochRIA	10%	$11.41^{\pm 0.046}_{-0.10}$	$8.44^{\pm 0.016}_{-0.04}$	$23.74^{\pm 0.230}_{+0.15}$	$26.78^{\pm 0.127}_{+0.65}$
4:8	RIA	Full	8.39	6.74	13.77	21.59
	stochRIA	10%	$8.44^{\pm 0.014}_{-0.05}$	$6.74^{\pm 0.013}_{+0}$	$13.93^{\pm 0.095}_{-0.16}$	$21.49^{\pm 0.089}_{+0.10}$

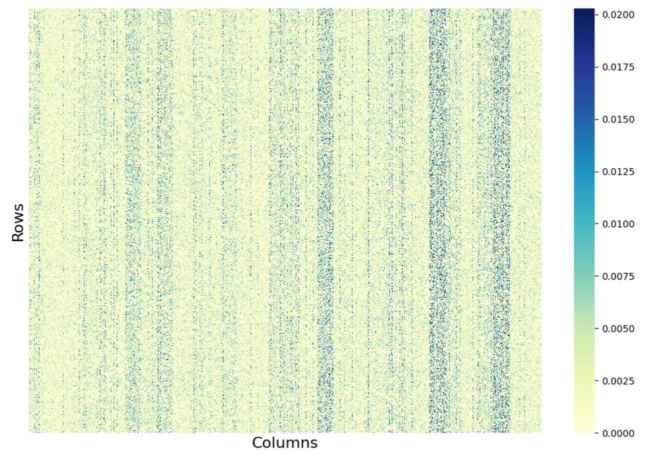

Figure 1. (a) Comparison of StochRIA ($\beta=0.1$) and RIA on Wikitext-2. Perplexity scores with $\alpha=1$ are reported. For StochRIA, the mean perplexity over 5 trials is shown in bold, with standard deviation in green. Improvements and declines relative to RIA are highlighted in blue and red, respectively. (b) Visualization of the dense weight matrix in LLaMA2-7b.

Table 2. Perplexity scores on Wikitext-2, accounting for various norm α values and column & row sensitivity, with a sparsity ratio 50%.

Model	LlaMA2-7b				LlaMA2-13b				LlaMA3-8b				OPT-1.3b			
α	0	0.5	1	2	0	0.5	1	2	0	0.5	1	2	0	0.5	1	2
Dense	5.47				4.88				6.14				14.62			
Wanda	16.03	7.60	7.79	8.66	6.83	6.17	6.28	7.15	205.44	10.66	10.81	12.98	1712.39	22.14	22.19	24.74
Col-Sum	11.59	6.83	6.91	7.46	6.39	**5.87**	5.96	6.55	59.41	9.53	9.69	12.01	1062.66	18.28	18.41	22.25
Row-Sum	14.93	7.49	7.51	8.01	6.74	6.13	6.24	7.01	17.80	10.50	10.55	11.79	141.92	22.09	22.47	26.62
RIA	7.39	**6.81**	6.88	7.37	5.95	5.93	5.95	6.56	12.07	**9.34**	9.44	10.67	64.70	**18.08**	18.94	23.39

(a) Perplexity on Wikitext-2 with different sparsity ϵ. $\alpha=1.0$.

Sparsity	Method	Sampling	L2-7b	L2-13b	L3-8b	OPT-1.3b
Dense	-	-	5.47	4.88	6.14	14.62
50%	Wanda	-	7.79	6.28	10.81	22.19
	RIA	Full	**6.88**	**5.95**	**9.44**	18.94
	stochRIA	10%	6.91	**5.95**	9.46	**18.78**
60%	Wanda	-	15.30	9.63	27.55	38.81
	RIA	Full	**10.39**	**7.84**	19.52	26.22
	stochRIA	10%	10.62	7.97	**19.04**	**25.93**
70%	Wanda	-	214.93	104.97	412.90	231.15
	RIA	Full	**68.75**	**51.96**	169.51	98.52
	stochRIA	10%	72.85	62.15	**155.34**	**93.29**

(b) Perplexity on Wikitext-2 after training-free fine-tuning. $\epsilon=60\%$ and $\alpha=0.5$.

Base	FT	LlaMA2-7b	LlaMA2-13b	LlaMA3-8b
Dense	-	5.47	4.88	6.14
Magnitude	-	6.9e3	10.10	4.05e5
Magnitude	DSnoT	4.1e3	10.19	4.18e4
Magnitude	R^2-DSnoT	**2.4e2**	**10.09**	**1.44e4**
Wanda	-	**9.72**	7.75	21.36
Wanda	DSnoT	10.23	**7.69**	20.70
Wanda	R^2-DSnoT	10.08	**7.69**	**20.50**
RIA	-	10.29	7.85	21.09
RIA	DSnoT	9.97	7.82	19.51
RIA	R^2-DSnoT	**9.96**	**7.78**	**18.99**

Figure 2. Perplexity results on Wikitext-2. (a) shows the performance under different sparsity levels, while (b) presents results after training-free fine-tuning.

Table 3. Accuracies (%) for LLaMA2 models on 7 zero-shot tasks at 60% unstructured sparsity.

Params	Method	BoolQ	RTE	HellaSwag	WinoGrande	ARC-e	ARC-c	OBQA	Mean
LlaMA2-7b	Dense	77.7	62.8	57.2	69.2	76.4	43.4	31.4	57.9
	Magnitude	41.2	51.3	37.0	55.7	50.0	27.0	16.2	39.3
	w. DSnoT	43.2	54.2	38.4	56.4	53.3	27.7	20.6	41.1
	w. R^2-DSnoT	50.9	52.0	39.8	56.8	56.6	28.3	23.4	**43.4**
	RIA	66.1	53.1	43.5	63.2	64.6	30.2	26.0	49.5
	w. DSnoT	65.5	53.4	44.7	64.6	65.3	31.7	26.4	50.2
	w. R^2-DSnoT	65.2	53.8	44.7	65.1	65.0	31.6	27.0	**50.3**
LlaMA3-8b	Dense	81.3	69.7	60.1	73.0	80.1	50.4	34.8	64.2
	Magnitude	37.8	52.7	30.7	51.0	39.7	23.4	14.4	35.7
	w. DSnoT	37.8	52.7	33.4	49.9	43.5	23.0	14.8	36.4
	w. R^2-DSnoT	37.8	52.7	33.1	52.1	43.9	23.6	14.8	**37.1**
	RIA	70.2	53.4	39.7	61.7	61.1	28.6	20.4	47.9
	w. DSnoT	70.7	53.4	40.3	61.3	61.7	28.0	20.0	47.9
	w. R^2-DSnoT	70.4	53.4	40.3	61.9	61.2	28.3	21.0	**48.1**

This study analyzed PTP methods, focusing on Wanda and RIA, offering empirical and theoretical insights into input activations and weight importance via the symmetric objective in (Sym). We introduced a training-free fine-tuning step within a prune-and-grow framework, outperforming baselines. These findings enhance the understanding of PTP and support future research on efficient LLM compression.