International Conference On Machine Learning

BaWA: Automatic Optimizing Pruning Metric for Large Language Models with Balanced Weight and Activation

Weight Normalization + Activation Regulation = Optimized Pruning Metric

Lian Liu, Xiandong Zhao, Guanchen Li, Dong Li, Mengdi Wang, Yinhe Han, Xiaowei Li, Ying Wang

Institute of Computing Technology, CAS

University of Chinese Academy of Sciences

Background: LLM Pruning Challenge

Large Language Models

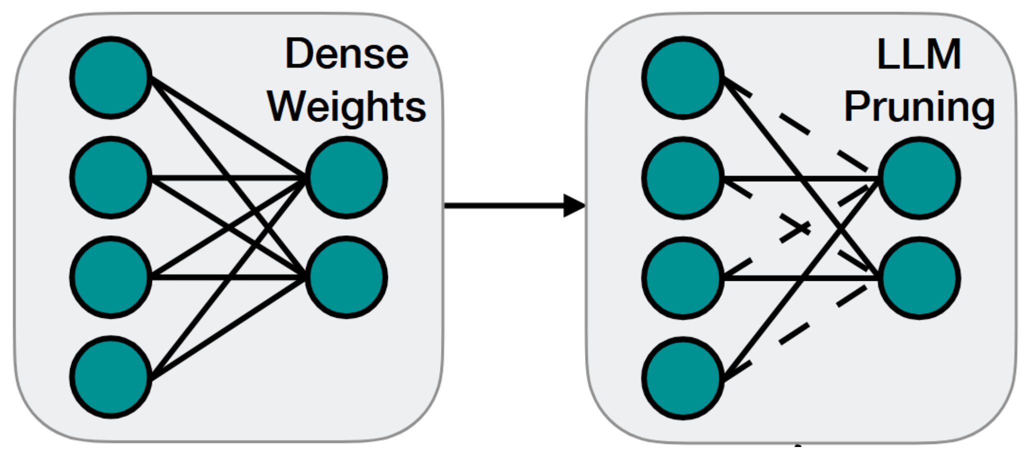
- Billions of parameters (e.g., LLaMA, Mistral, Qwen2)
- Exceptional capabilities across diverse tasks
- Significant hardware constraints for deployment

Model Compression Challenge

High Memory	High Compute	Slow Inference
40+ GB	**175B+ params**	**Limited throughput**

Pruning Solutions

- Removes redundant weights to reduce model size
- **One-shot post-training pruning:** Efficient approach without fine-tuning
- Can achieve 50%+ sparsity with minimal performance loss
- Supported by hardware acceleration (e.g., 2:4 sparse tensor cores)

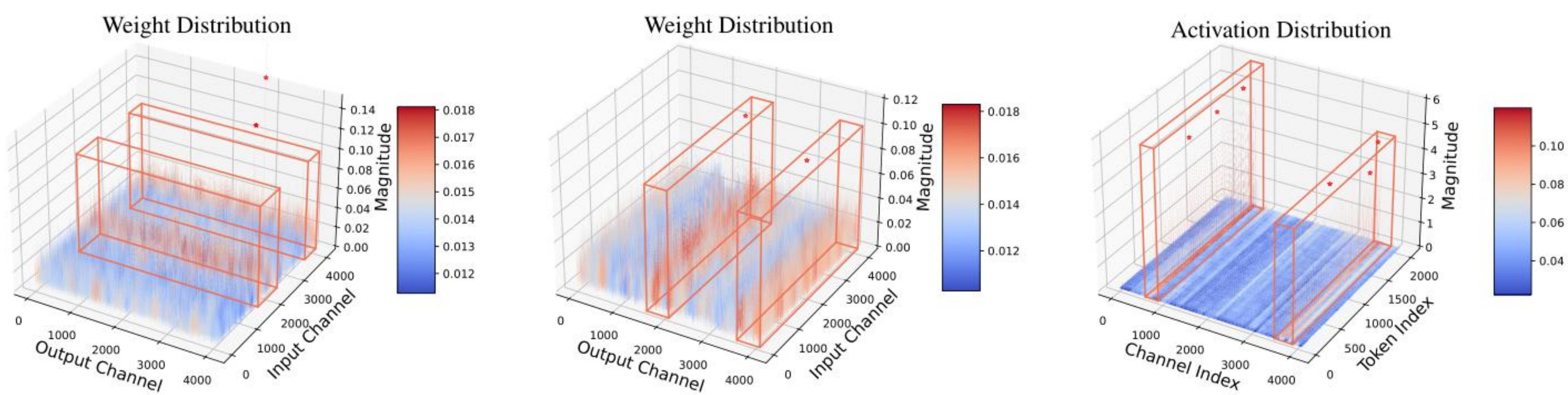


Limitations of Current Pruning Methods

Current Pruning Metrics: Simple Symbolic Combinations

Magnitude	Wanda	GBLM-Pruner
$\|Wij\|$	$\|Wij\| \cdot \|\|Xj\|\|2$	$\|Wij\| \cdot \|\|Xj\|\|2 \cdot \|Gij\|$

These methods overlook critical distribution characteristics in weights and activations

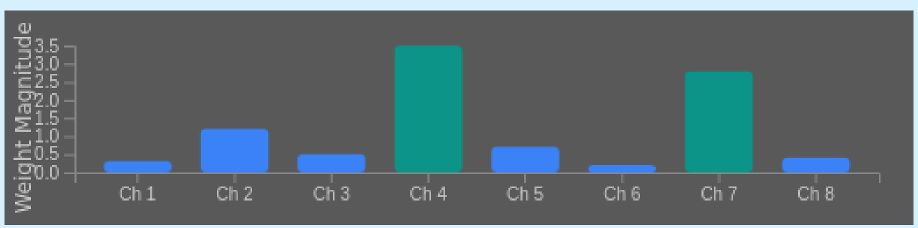


(a) layers.10.self_attn.o_proj (b) layers.10.self_attn.v_proj (c) layers.10.self_attn.k_proj

Key Observations: Why Current Methods Fail

Imbalanced Weight Magnitude Distribution

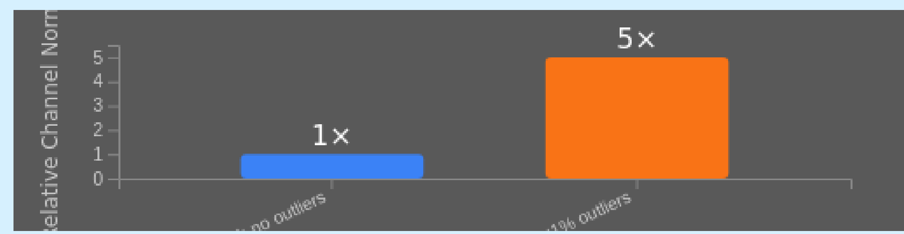
- Weight magnitudes vary significantly across channels
- Certain channels contain abnormally large or small weights
- Leads to biased pruning decisions where entire channels are either preserved or pruned



Weight magnitude varies significantly across different channels

Disproportionate Impact of Outliers

- Less than 1% of activation outliers can inflate channel's norm by up to 5×
- Channels with outliers are erroneously prioritized during pruning
- Channels without outliers are excessively pruned, degrading performance



A few outliers dramatically increase channel norm values

Key insight: Current pruning metrics use simple symbolic combinations of weights and activations, ignoring these imbalances. This leads to sub-optimal pruning decisions and significant performance degradation.

Introducing BaWA

BaWA (Balanced Weight and Activation) is a novel pruning metric that systematically balances the contributions of weight and activation distributions for more effective LLM pruning, addressing the limitations of existing methods

BaWA Pruning Process

Imbalanced Distribution → Magnitude Normalization → Outlier Regularization → Automatic Optimization

Magnitude Normalization

Normalizes weight magnitudes across both input and output channels to address imbalanced weight distributions, contributing to fairer pruning decisions.

Outlier Regularization

Introduces learnable power factors to reduce the impact of activation outliers, preventing their disproportionate influence on pruning decisions.

Automatic Optimization

Employs zeroth-order gradient optimization to efficiently search for optimal hyperparameters, enabling better pruning masks with minimal computational overhead.

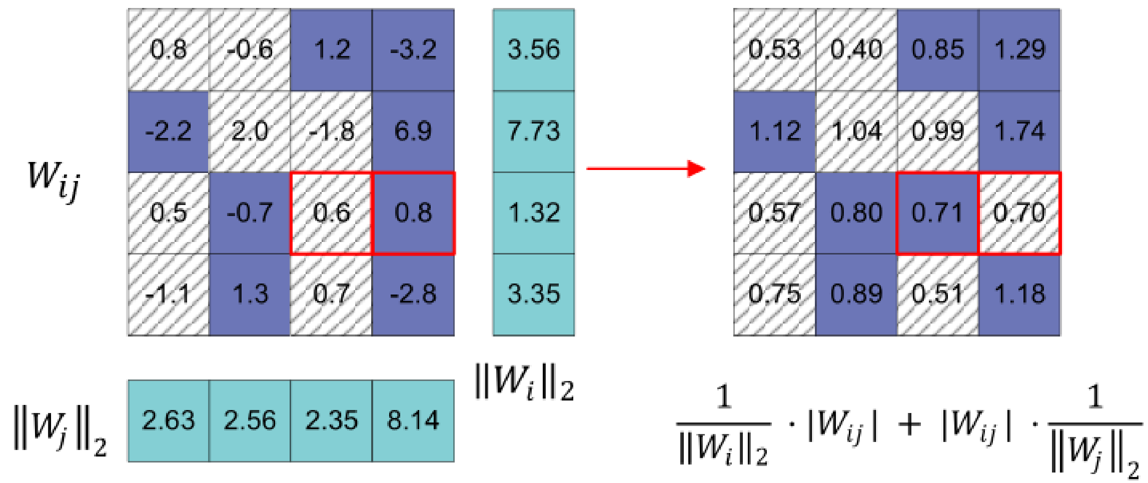
Magnitude Normalization

The Problem

Weight magnitudes exhibit significant imbalance across channels

Some channels contain weights that are abnormally large or small

This leads to biased pruning where weights in certain channels are predominantly preserved or removed



(a) Magnitude Normalization

BaWA's Solution

Input Channel Normalization

Normalizes weight magnitude by the ℓ2-norm of each input channel

$S_{ij}^{(ICN)} = |W_{ij}| \cdot (1/\|W_j\|_2) \cdot \|X_j\|_2$

Output Channel Normalization

Normalizes by the ℓ2-norm of each output channel

$S_{ij}^{(OCN)} = (1/\|W_i\|_2) \cdot |W_{ij}| \cdot \|X_j\|_2$

Benefits of Magnitude Normalization

- More balanced distribution
- Fairer pruning decisions
- Improved model performance
- Optimal sparsity patterns

Outlier Regularization

The Outlier Problem

Few activation outliers (<1%) can inflate a channel's norm by over 5×

Channels without outliers are unfairly pruned

Up to 10% of channels eliminated in specific layers

Existing metrics over-emphasize outlier channels

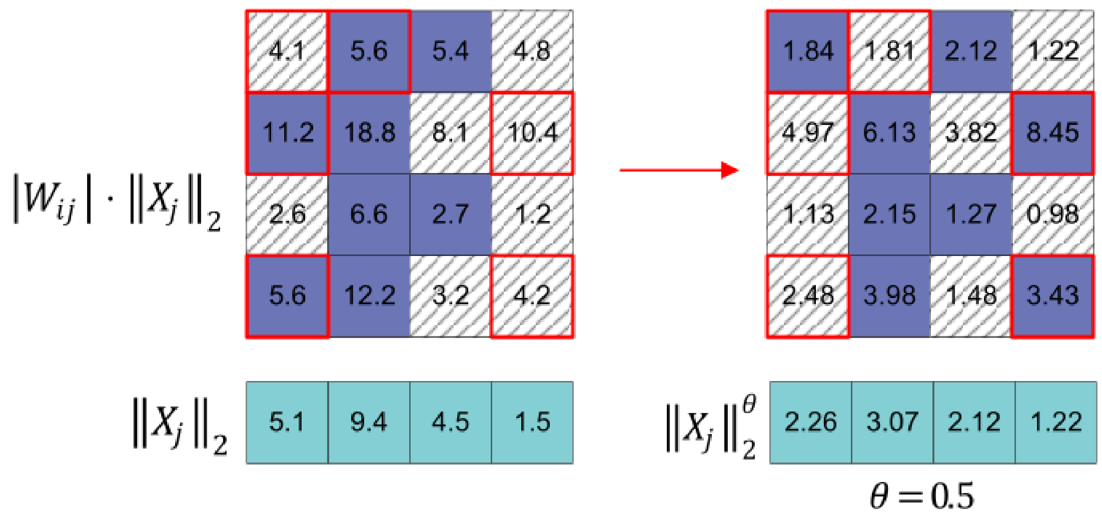
BaWA's Solution

Introduce power factor θ to control outlier influence

Lower θ values reduce impact of activation outliers

Learnable parameters optimize regularization strength

Ensures fair evaluation of each weight's importance



(b) Outlier Regularization

Automatic Hyperparameter Optimization

$$\boldsymbol{S}_{ij} = \Big(\underbrace{|\boldsymbol{W}_{ij}| \cdot \frac{1}{||\boldsymbol{W}_j||_2^{\theta_1}}}_{\text{input channel normalization}} + \underbrace{\frac{1}{||\boldsymbol{W}_i||_2^{\theta_2}} \cdot |\boldsymbol{W}_{ij}|}_{\text{output channel normalization}} \Big) \cdot ||\boldsymbol{X}_j||_2^{\theta_3}$$

$$\begin{aligned} \Theta^* &= \arg\min_{\Theta} \mathcal{L}(\Theta; \boldsymbol{X}), \\ \mathcal{L}(\Theta; \boldsymbol{X}) &= \|\mathrm{RMSNorm}\left(\mathcal{F}(\mathbb{W}; \boldsymbol{X})\right) \\ &\qquad - \mathrm{RMSNorm}\left(\mathcal{F}(\mathbb{W} \odot \mathbb{M}; \boldsymbol{X})\right)\|_2^2, \\ \mathbb{M} &= \mathbb{S} > \mathrm{top}_{\mathrm{k}}(\mathbb{S}), \end{aligned}$$

Optimization Process

Initialize Parameters

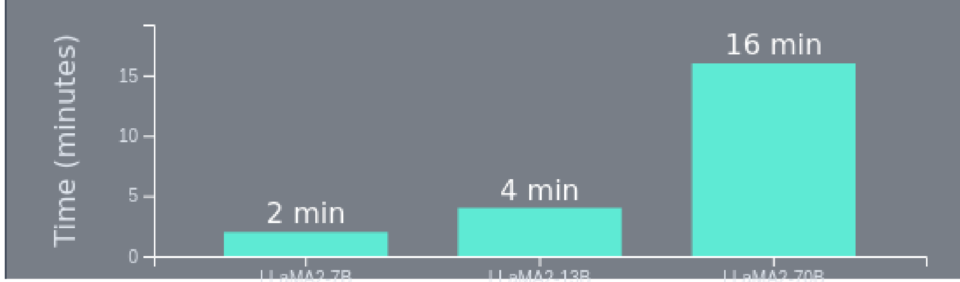
Estimate Gradients

Update Parameters

Block-wise optimization reduces search complexity

Uses calibration dataset (128 segments from C4)

Optimization Efficiency



Experimental Results: Perplexity

WikiText-2 perplexity performance of BaWA and Wanda for different LLMs at varying sparsity rates.

	LLaMA-7B			LLaMA-13B			LLaMA2-70B			Qwen2-72B		
Sparsity	60%	70%	80%	60%	70%	80%	60%	70%	80%	60%	70%	80%
Wanda	10.57	74.79	4.80e3	8.69	51.94	4.95e3	4.97	10.23	149.76	6.26	9.00	40.50
BaWA	**10.00**	**57.84**	**3.95e3**	**7.67**	**33.83**	**4.10e3**	**4.56**	**8.71**	**125.71**	**6.03**	**8.17**	**31.89**

		LLaMA2		Mistral	Qwen2
Method	Sparsity	13B	70B	7B	72B
Dense	0%	4.57	3.12	5.25	4.94
Magnitude	4:8	6.76	5.54	9.21	8.14
SparseGPT	4:8	6.60	4.59	8.07	5.97
Wanda	4:8	6.55	4.47	8.41	5.86
GBLM	4:8	6.54	4.49	8.31	5.85
RIA	4:8	6.29	4.37	8.27	5.81
Pruner-Zero	4:8	6.75	4.45	8.11	5.85
DSnoT	4:8	6.43	4.41	7.93	5.79
ADMM-Iter	4:8	6.37	4.35	7.79	5.77
BaWA	4:8	6.16	4.32	7.54	5.74
BaWA + ADMM	4:8	**6.07**	**4.24**	**7.36**	**5.65**

⭐ Key Improvements

- **LLaMA-7B (60%):** 0.57 perplexity reduction vs. Wanda
- **LLaMA-13B (70%):** 18.11 perplexity reduction vs. Wanda
- **Qwen2-72B (80%):** 8.61 perplexity reduction vs. Wanda
- **LLaMA2-70B (4:8):** 0.15 perplexity reduction vs. Wanda

BaWA consistently outperforms all baseline methods across various models and sparsity levels

Experimental Results: Zero-Shot Tasks

			LLaMA				LLaMA2				
Method	Weight Update	Sparsity	7B	13B	30B	65B	7B	13B	70B	Mistral-7B	Qwen2-72B
Dense	-	0%	59.99	62.59	65.38	66.97	59.71	63.03	67.08	64.30	69.82
Magnitude	✗	50%	46.94	47.61	53.83	62.74	51.14	52.77	60.93	55.87	60.66
SparseGPT	✓	50%	54.94	58.61	63.09	66.30	56.24	60.57	67.28	59.34	68.11
Wanda	✗	50%	55.13	59.33	63.60	66.67	56.24	60.04	67.03	58.93	66.41
BaWA	✗	50%	**55.27**	**59.97**	**64.12**	**67.21**	**57.02**	**60.67**	**67.81**	**60.17**	**69.11**

Consistent Improvement

BaWA outperforms Wanda by **up to 3.08%** on average accuracy across tasks

Superior Performance

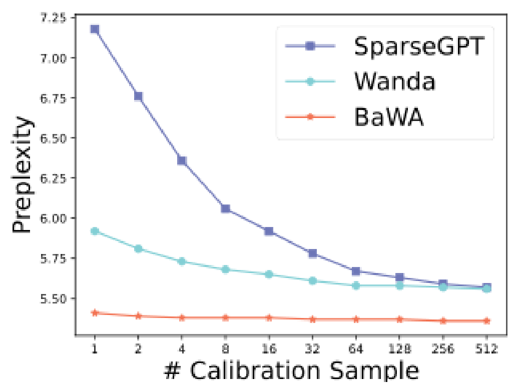
For Mistral-7B with 2:4 sparsity, BaWA shows **53.23%** accuracy vs. Wanda's 50.15%

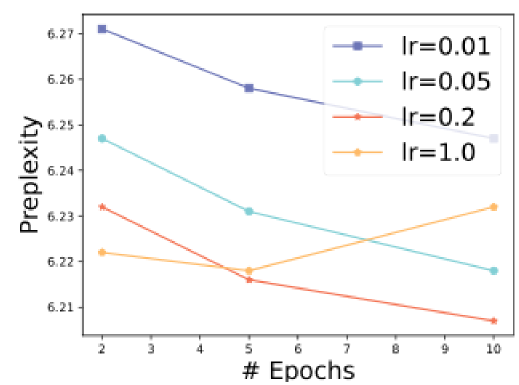
Model Adaptability

For LLaMA2-70B, the pruned model with 50% sparsity achieves **higher accuracy** than the original dense model

Experimental Results: Analysis

	LLaMA2 & Qwen2 (50%)			LLaMA2 & Qwen2 (4 : 8)			LLaMA2 & Qwen2 (2 : 4)		
Method	13B	70B	72B	13B	70B	72B	13B	70B	72B
Wanda	5.56	3.98	5.48	6.55	4.47	5.86	8.27	5.16	6.31
Input Channel Normalization	5.47	3.89	5.48	6.38	4.42	5.84	7.93	5.13	6.30
Magnitude Normalization	5.45	3.88	5.44	6.27	4.41	5.81	7.74	5.04	6.27
Outlier Regularization (0.5)	5.46	3.90	5.46	6.20	4.39	5.77	7.54	4.95	6.21
BaWA w/o Automatic Search	5.45	3.88	5.43	6.27	4.41	5.80	7.74	5.05	6.23
BaWA w/ Automatic Search	**5.42**	**3.84**	**5.41**	**6.16**	**4.32**	**5.74**	**7.13**	**4.84**	**6.14**





Ablation Study demonstrates the effectiveness of each method proposed by BaWA

Conclusion and Impact

Key Contributions

Balanced Pruning Metric
Addresses imbalanced weight magnitudes and disproportionate influence of activation outliers

Superior Performance
For Mistral-7B with 2:4 sparsity: reduced perplexity by 2.49 and improved downstream task accuracy by 3.08%

Efficient Implementation
Complete optimization in ~16 minutes for LLaMA2-70B on a single GPU, with minimal performance overhead

Impact & Significance

- Consistently outperforms existing SOTA pruning methods across various LLMs and language benchmarks
- Compatible with existing weight reconstruction methods (e.g., ADMM-Iter), offering further performance gains
- Enables effective deployment of LLMs in resource-constrained environments
- Orthogonal to conventional weight adjustment methods, creating opportunities for combined approaches

Performance Highlights

1.58×	3.08%	50%+
Speedup over dense FP16 GEMM	Improved accuracy on downstream tasks	Effective at high sparsity levels

Thank you for your attention!

Questions?