

DLP: Dynamic Layerwise Pruning in Large Language Models

动态层级剪枝

ICML 2025 第42届国际机器学习大会

Beijing University of Posts and Telecommunications

Hong Kong University of Science and Technology

Background

- 大语言模型的挑战

- LLaMA-2、LLaMA-3、GPT-3 等模型参数规模数十亿甚至千亿，推理耗时、存储和能耗成本过高。
- 在资源受限环境（如 CPU、边缘设备）下部署难度大。

- 现有剪枝方法的不足

1.统一稀疏率分配 (Uniform Pruning, 如 SparseGPT, Wanda)

- 每一层分配相同的剪枝比例。
- 忽视了不同层的重要性差异 → 在高稀疏率下性能严重退化。
- 例：某些关键层被过度剪枝，模型语义能力崩溃。

2.非均匀稀疏率分配 (如 OWL)

- 基于离群值比例来调整层级稀疏率。
- 存在两个问题：
- 需要设定经验阈值，缺乏自适应性。
- 稳定性不足，不同模型/数据集迁移性差。

Method: DLP (Dynamic Layerwise Pruning)

DLP 的核心思想：**根据层的重要性动态分配稀疏率**，而不是一刀切。

1. 层重要性度量

- 使用 **权重和输入激活** 来度量层级“不重要性”。
- 采用 **中位数 (median)** 作为统计量：

好处：对 LLM 中常见的 **outliers (离群值激活)** 更稳健。

如果用均值 (mean)，容易被极大值误导，导致错误判断层的重要性。

$$\mathbf{A}_{ij}^l = |\mathbf{W}_{ij}^l| \cdot \|\mathbf{X}_j^l\|_2$$

$$S^l = \sum_{i=1}^{C_{\text{out}}} \sum_{j=1}^{C_{\text{in}}} \mathbf{F}(\mathbf{A}_{ij}^l)$$

2. 相对重要性分布 (Relative Importance Distribution, RID)

将所有层的不重要性分数归一化，得到各层的 **相对重要性分布**。

$$I^l = 1 - \frac{S^l}{\sum_{i=1}^L S^i}$$

The importance scores of all layers constitute the RID, that is, $\text{RID} = [I^1, I^2, \dots, I^L]$. Layers with lower importance have less impact on model performance, so they should be assigned higher sparsity.

Method: DLP (Dynamic Layerwise Pruning)

DLP 的核心思想：根据层的重要性动态分配稀疏率，而不是一刀切。

3. 动态稀疏率分配策略

- 设置一个超参数 α ，用于约束每层稀疏率的范围，避免某些层被过度或不足剪枝。

Algorithm 1 Pseudocode of DLP

1: **Input:** Weight $\mathbf{W}$ and input $\mathbf{X}$

2: **Input:** Deflation scale α and sparsity rates p

3: **Output:** The dynamic pruning sparsity R of each layer

4: Obtain the score $\mathbf{A}_{ij}^l$ via Eq.(3)

5: Obtain the unimportance score S^l via Eq.(5)

6: Obtain the importance score I^l via Eq.(6)

7: Store the scaled importance scores in d

8: **for** $i \leftarrow 0$ **to** $\text{length}(I) - 1$ **do**

9: $d_j \leftarrow \frac{I_i - I_{\min}}{I_{\max} - I_{\min}} \times 2 \times \alpha$

10: **end for**

11: Calculate the average values m of d

12: **for** $j \leftarrow 0$ **to** $\text{length}(d) - 1$ **do**

13: $R_j \leftarrow p + m - d_j$

14: **end for**

15: Obtain the final layerwise sparsity rates R

16: **Return** R

缩放每层的重要性，得到动态调节因子 d_j
归一化到 $[0, 1]$ ，再缩放到 $[0, 2\alpha]$
表示第 j 层稀疏率要比平均值高/低多少

分配每层的最终稀疏率
如果某层重要性高 $\rightarrow d_j$ 大 $\rightarrow R_j$ 小（剪得少）
如果某层不重要 $\rightarrow d_j$ 小 $\rightarrow R_j$ 大（剪得多）

计算所有层的平均偏移量 m

保证整体平均稀疏率仍等于 p

Experiments

Table 3. Perplexity results on WikiText. We produce the Uniform, OWL and DLP(Ours) with 70% unstructured sparsity on LLaMA1, LLaMA2 models. The best performance result is indicated in bold.

Method	Layerwise Sparsity	LLaMA1			LLaMA2	
		7B	13B	30B	7B	13B
Dense	-	5.68	5.09	4.10	5.47	4.88
Magnitude	Uniform	4.9e4	8.5e4	9.7e2	5.0e4	2.1e2
	OWL	2.0e4	1.9e4	2.4e2	1.5e4	57.55
	Ours	3.4e3	7.6e3	98.05	8.7e3	52.41
SparseGPT	Uniform	25.38	18.93	12.87	27.84	19.38
	OWL	19.95	14.02	10.22	19.71	15.12
	Ours	17.76	12.63	9.43	18.58	13.30
Wanda	Uniform	86.38	56.26	17.54	76.84	45.76
	OWL	24.46	16.23	10.77	30.58	20.65
	Ours	20.46	13.65	9.93	22.79	16.19

Table 4. Comparison of mean zero-shot accuracies (%) for pruned LLaMA1 and LLaMA2 models at 70% unstructured sparsity. The best performance result is indicated in bold.

Method	Layerwise Sparsity	LLaMA1			LLaMA2	
		7B	13B	30B	7B	13B
Dense	-	64.33	66.78	69.72	64.42	67.04
Magnitude	Uniform	34.80	37.09	35.06	35.65	36.23
	OWL	36.40	39.45	35.73	36.44	40.67
	Ours	38.21	40.11	42.03	40.84	44.02
SparseGPT	Uniform	45.32	48.34	55.87	44.72	47.99
	OWL	47.84	50.78	56.82	48.02	51.70
	Ours	48.32	53.06	57.84	49.65	53.47
Wanda	Uniform	39.91	41.62	54.59	37.04	40.44
	OWL	46.32	49.59	55.93	43.55	48.11
	Ours	48.62	52.03	56.83	46.25	51.11

Experiments

Table 5. End-to-end decoding latency and throughput of LLaMA2-7B-chat-hf on DeepSparse inference engine using DLP.

Sparsity	Dense	10%	20%	30%	40%	50%	60%	70%	80%	90%
Latency (ms)	353.42	352.40	338.37	323.49	273.43	200.17	164.31	124.76	100.30	96.86
Throughput (tokens/sec)	2.83	2.84	2.96	3.09	3.66	4.99	6.08	8.01	9.97	10.32
Speedup	1.0x	1.0x	1.1x	1.1x	1.3x	1.8x	2.2x	2.8x	3.5x	3.7x

Experiments

Table 7. WikiText validation perplexity of various LLMs pruned by Ours with LoRA fine-tuning.

Model	Method	Sparsity	Perplexity
LLaMA1-7B	Without FT	0.7	17.76
LLaMA1-7B	With FT	0.7	12.15
LLaMA1-13B	Without FT	0.7	12.63
LLaMA1-13B	With FT	0.7	10.05

Experiments

Table 8. WikiText validation perplexity of various LLMs pruned by Uniform and Ours using Wanda. The best performance result is indicated in bold.

Model	Method	60%	70%	80%
LLaMA3-8B	Uniform	23.50	122.96	687.11
	Ours	19.21	96.31	676.16
LLaMA3.1-8B	Uniform	21.84	118.18	1031.36
	Ours	18.58	84.30	786.19
Vicuna-7B	Uniform	12.89	60.60	1613.15
	Ours	11.49	28.79	345.10
Mistral-7B	Uniform	11.28	60.62	331.04
	Ours	9.91	29.83	199.19
Qwen-7B	Uniform	14.76	91.99	23136.98
	Ours	14.38	54.13	1122.54