Distribution-aware Pruning Strategy for Large Language Models[1]

Qi Lei

with Jianwei Li and Yijun Dong

Courant Institute & Center for Data Science, NYU

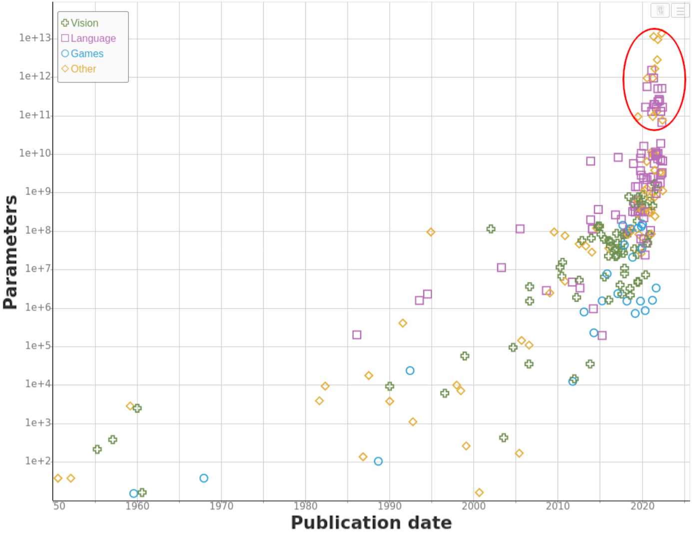
JSM, 2024.08

[1] https://arxiv.org/pdf/2407.19126

Contents

Introduction

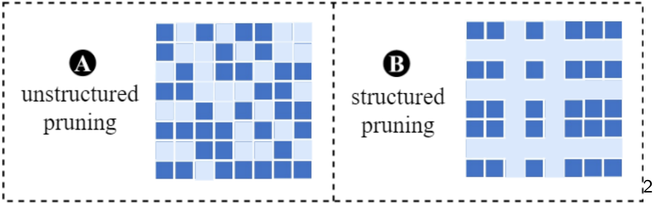
Motivation



[Villalobos et al. 2022]

Prior work

Classification 1: model structure preservation



[2]

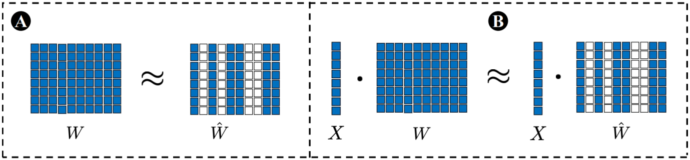
A: better performance preservation

B: hardware compatibility; efficient at inference time

[2] Pruning masks: Dark blue is kept weight; light blue is pruned out weight.

Prior work

Classification 2: approximation principle



A: Preserving model weights

B: Preserving model outputs

Prior work

Classification 3: Retraining requirements (Computational costs)

A: Iterative pruning (High)

B: Finetuning-required pruning (Median)

C: One-shot pruning (Relatively Low)

Value-based $\ll$ Gradient-based $\ll$ Hessian-based

Methodology

Goal

Iterative pruning	==>	Single-shot pruning
Unstructured pruning	==>	Structured pruning
Gradient/Hessian-based	==>	Value-based pruning
Weight preservation	==>	Output preservation

Goal

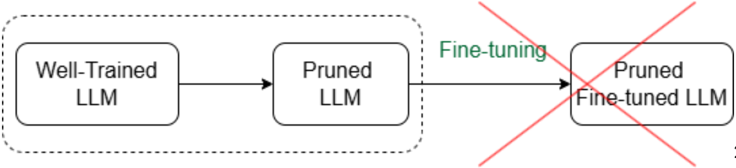
Iterative pruning	==>	Single-shot pruning
Unstructured pruning	==>	Structured pruning
Gradient/Hessian-based	==>	Value-based pruning
Weight preservation	==>	Output preservation

One-shot Pruning

[2]

[2] Concentrate on the effectiveness of the pruning method, instead of comparisons of fine-tuning data's quality.

One-shot Pruning



[2]

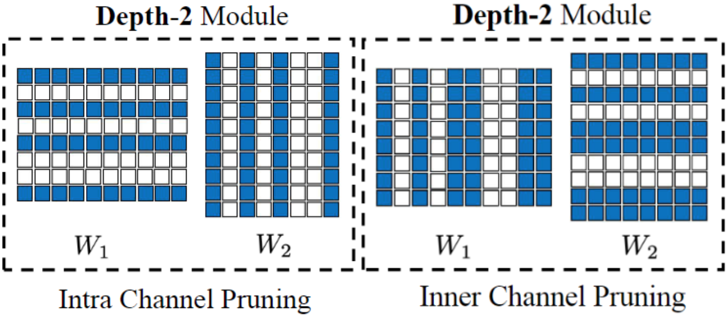
[2] Concentrate on the effectiveness of the pruning method, instead of comparisons of fine-tuning data's quality.

Goal

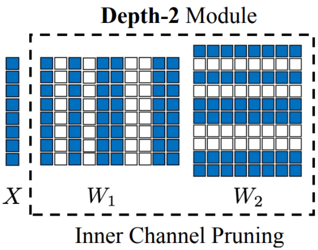
Iterative pruning	==>	Single-shot pruning
Unstructured pruning	==>	Structured pruning
Gradient/Hessian-based	==>	Value-based pruning
Weight preservation	==>	Output preservation

Pruning Unit: Depth-2 Units

Two pruning strategies:



Depth-2 Unit 1: Feedforward Layer



Depth-1 magnitude-based pruning: $\|(W_1)_{:,i}\|$

Depth-2 magnitude-based pruning: $\|(W_1)_{:,i}\|\|(W_2)_{i,:}\|$

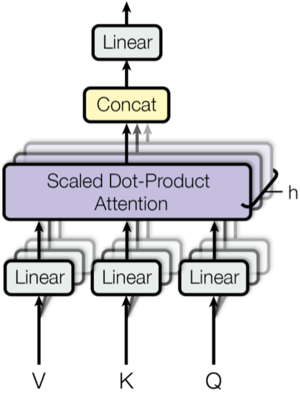
Ours: $\|(W_2)_{i,:}\|^2 (W_1)_{:,i}^\top \Sigma (W_1)_{:,i}$

Rational: magnitude of each slice $E[\|(W_2)_{i,:}\|^2 \sigma^2((W_1)_{:,i}^\top X)]$

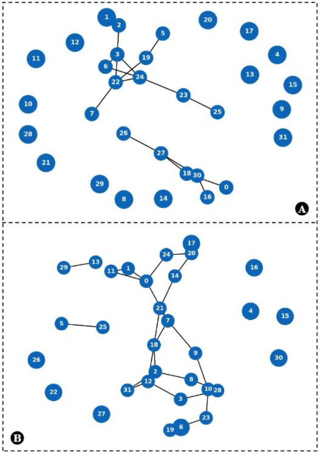
$$= \frac{1}{2}\|(W_2)_{i,:}\|^2 (W_1)_{:,i}^\top \Sigma (W_1)_{:,i}.$$

(Take input X as a normal distribution with covariance Σ, σ is ReLU.)

Depth-2 Unit 2: Attention Layer



multi-head attention



32 attention heads from Block 4&5 of Llama-7B
Connected if $D(h_i, h_j) \geq 0.2$.

Goal

Iterative pruning	==>	Single-shot pruning
Unstructured pruning	==>	Structured pruning
Gradient/Hessian-based	==>	Value-based pruning
Weight preservation	==>	Output preservation

Layer-wise Recovery

Motivation:

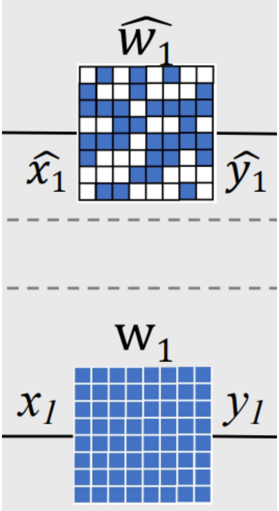
- For gradient-based pruning ==> global criterion ==> $f(\cdot; W + \Delta W) \approx f(\cdot; W) + \nabla_W f(\cdot, W)\Delta W$
- For Value-based pruning ==> local criterion for each layer ==> error will compound layer by layer (if each layer is pruned independently)

Layer-wise Recovery from Targeted Value

We will apply the above pruning strategy on a recovered weight $\hat{W}_l$:

$$\hat{W}_l \leftarrow \arg\min_{W} \|W\hat{X}_l - Y_l\|,$$

$\hat{X}_l$ is the updated input due to pruned weights $\hat{W}_1, \cdots \hat{W}_{l-1}$, Y_l is the targeted output. [a]



[a] [Li, L, Cheng, Xu, 2023]
https://arxiv.org/abs/2310.13191

Results

Results

Methods	WikiText2↓	PTB↓	BoolQ	PIQA	HS	WG	ARC-e	ARC-c	OBQA	Ave ↑
Dense	12.62	22.14	73.18	78.35	72.99	67.01	67.45	41.38	42.4	63.5
Data Free Pruning										
Random	23.02	40.19	46.21	71.33	59.35	56.51	47.97	32.0	36.30	49.95
L1 norm	179.02	311.75	51.28	60.22	43.14	52.01	36.53	27.89	30.8	43.12
L2 norm	582.41	1022.17	60.18	58.54	37.04	53.27	32.91	27.56	29.8	42.76
Ours (Self-Gen)	21.76	34.3	63.51	72.63	56.54	54.46	51.68	33.79	36.4	52.72
Ours SG w/ remedy	**20.32**	**33.42**	**64.17**	**72.67**	**58.43**	**57.29**	**53.32**	**34.15**	**37.23**	**53.89**
Data Dependent Pruning										
Training-Aware Pruning										
LLM-Pruner Vec	22.28	41.78	61.44	71.71	57.27	54.22	55.77	33.96	38.4	53.52
LLM-Pruner E1	19.09	34.21	57.06	75.68	66.8	59.83	60.94	36.52	40.0	56.69
LLM-Pruner E2	19.77	36.66	59.39	75.57	65.34	61.33	59.18	37.12	39.8	56.82
Inference-Aware Pruning										
Wanda-sp	27.45	49.52	64.16	75.21	**68.62**	62.27	59.68	36.68	39.2	57.97
Ours (Calibration)	**17.48**	**30.04**	66.48	75.78	67.73	62.27	61.4	35.49	39.6	58.39
Ours C w/ remedy	17.90	31.23	**70.12**	**76.86**	68.55	**65.76**	**64.23**	**38.54**	**40.5**	**60.65**
Retraining-required Pruning										
LLM-P. LoRA	17.37	30.39	69.54	76.44	68.11	65.11	63.43	37.88	40.0	60.07

Model: LLaMA-7B (20% sparsity)
First two datasets: zero-shot perplexity (PPL) analysis
Next 7 datasets: zero-shot task classification

Conclusion

Conclusion

- Identifying inherent pruning structure:
  depth-2 units & attention heads
- Designing effective pruning criterion:
  distribution-aware value-based pruning
- Low-computational performance recovery technique:
  avoid error compound

Thank you!