第十章 大模型轻量化部署

修贤超

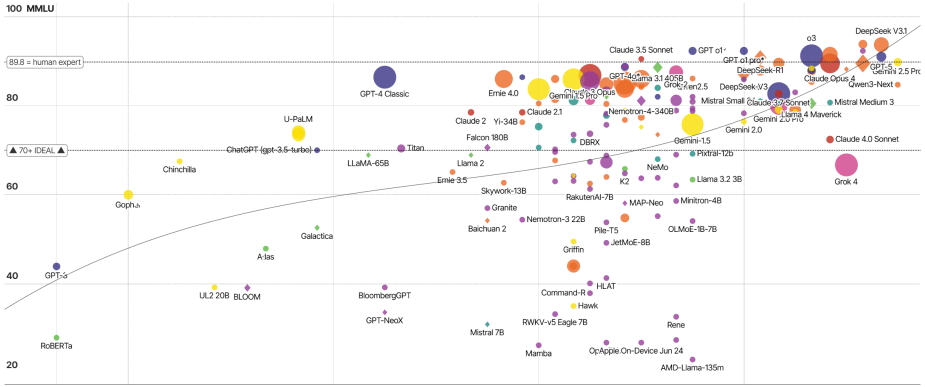
https://xianchaoxiu.github.io

致谢：本教案由徐力协助准备

目录

背景

- 大模型的规模与性能持续增长

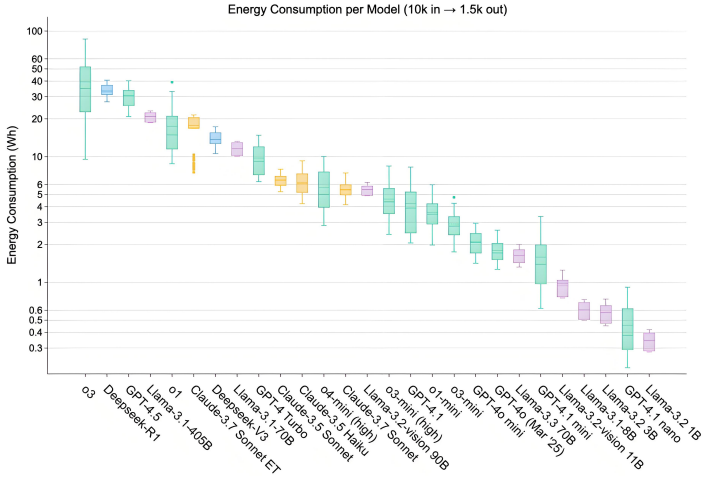


David McCandless, Tom Evans, Paul Barton
Informationisbeautiful // Sep 2025

MMLU = benchmark for measuring LLM capabilities
* = parameters undisclosed // source: LifeArchitect // data

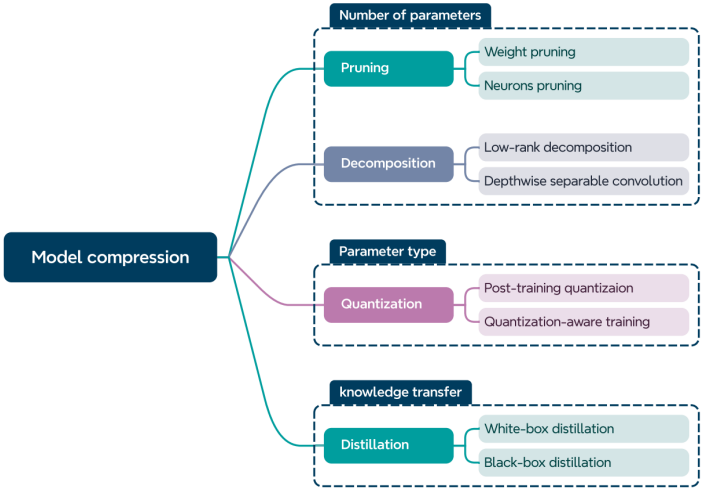
背景

- 模型参数增长带来一系列常见问题，比如推理延迟、显存消耗、能耗过高



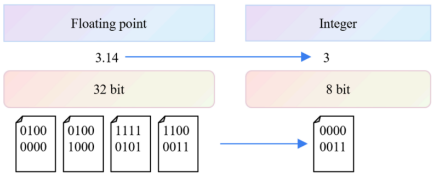
背景

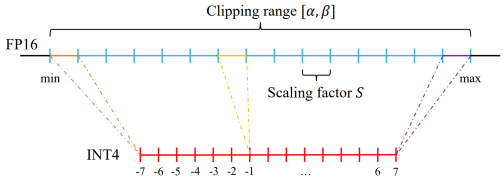
- A Survey of Model Compression Techniques: Past, Present, and Future, 2025



量化

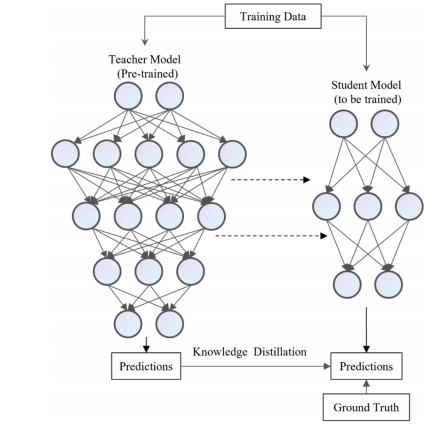
- 将模型权重从高精度浮点数转为低精度整数，这样每个参数占用更少字节，从而模型更小和推理更快

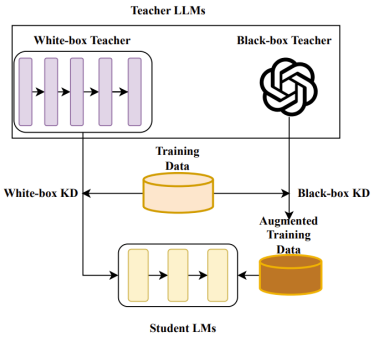




蒸馏

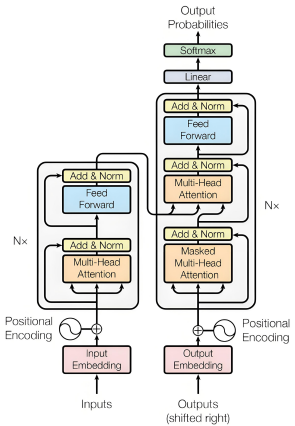
- 让一个小模型学习大模型的知识，使其在模型规模更小的情况下，依然具备与大模型相近的推理能力和泛化能力

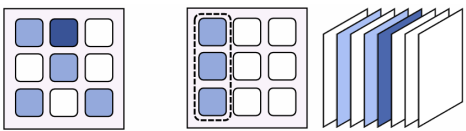




剪枝

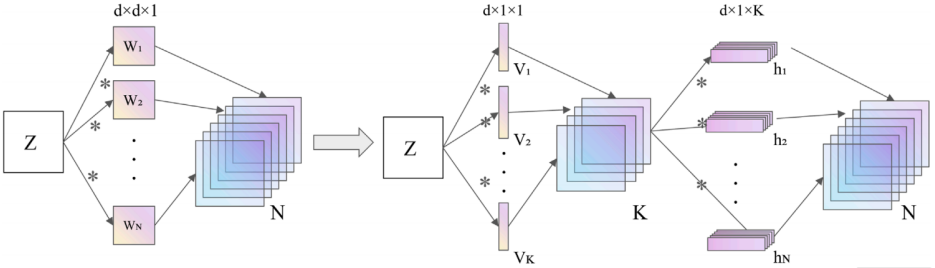
- 删除模型中影响较小的参数或结构，减少计算量和存储需求





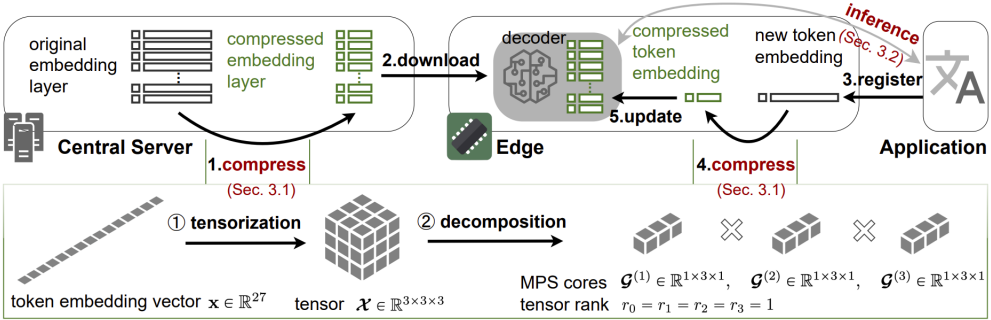
低秩分解

- 将大型矩阵分解为若干低秩矩阵乘积，可以减少参数量和计算量，同时尽量保持模型性能



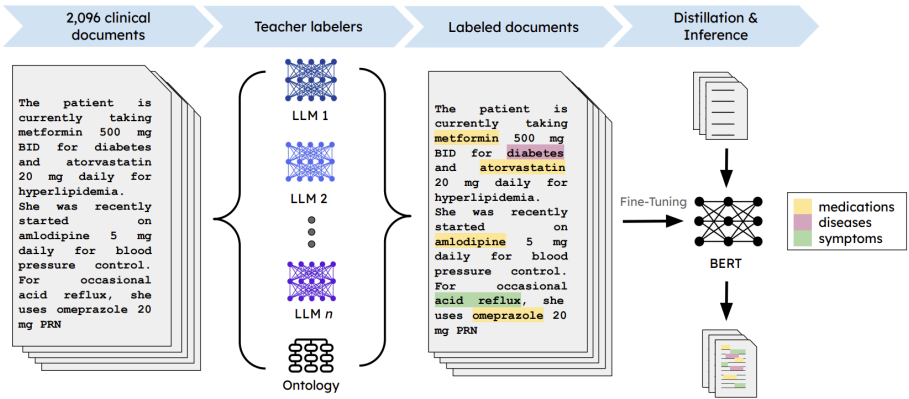
应用

- TensorSLM: Energy-efficient Embedding Compression of Sub-billion Parameter Language Models on Low-end Devices, 2025



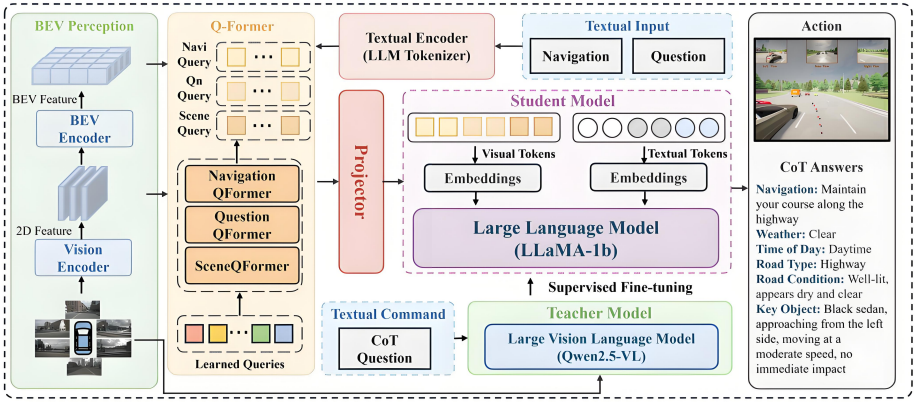
应用

- Distilling Large Language Models for Efficient Clinical Information Extraction, 2025



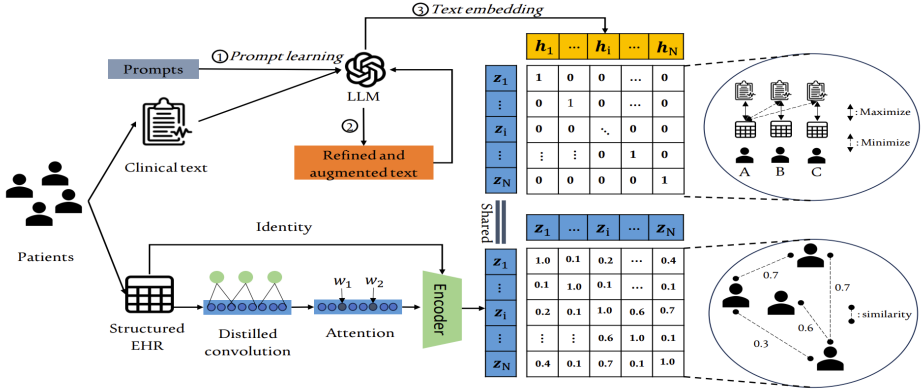
应用

- DSDrive: Distilling Large Language Model for Lightweight End-to-End Autonomous Driving, 2025



应用

- Distilling the Knowledge from Large-Language Model for Health Event Prediction, 2024



目录

非结构化剪枝

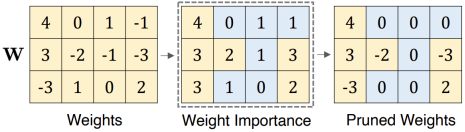
- Magnitude

Learning both weights and connections for efficient neural network

S Han, J Pool, J Tran, W Dally - Advances in neural …, 2015 - proceedings.neurips.cc

Neural networks are both computationally intensive and memory intensive, making them difficult to deploy on embedded systems. Also, conventional networks fix the architecture …

☆ 保存 99 引用 被引用次数：9600 相关文章 »



非结构化剪枝

- SparseGPT

  Sparsegpt: Massive language models can be accurately pruned in one-shot
  E Frantar, D Alistarh - International conference on machine …, 2023 - proceedings.mlr.press
  We show for the first time that large-scale generative pretrained transformer (GPT) family models can be pruned to at least 50% sparsity in one-shot, without any retraining, at minimal …
  ☆ 保存 99 引用 被引用次数：1018 相关文章 »

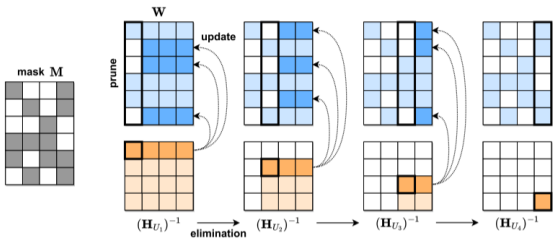
- 问题建模

$$\min_{M_\ell, \hat{W}_\ell} \|W_\ell X_\ell - (M_\ell \odot \hat{W}_\ell) X_\ell\|_2^2$$

非结构化剪枝

- Optimal Brain Surgeon（OBS）原理

$$\delta_m = -\frac{w_m}{[H^{-1}]_{mm}} \cdot H^{-1}_{:m}, \quad \varepsilon_m = \frac{w_m^2}{[H]^{-1}_{mm}}$$

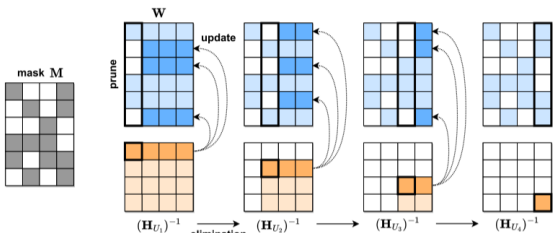


非结构化剪枝

- 共用 Hessian

$$U_{j+1} = U_j - \{j\}\ with\ U_1 = \{1, ..., d_{col}\}$$

$$(H_{U_{j+1}})^{-1} = \left(B - \frac{1}{[B]_{11}} \cdot B_{:1} B_{1:} \right)_{2:,2:}$$



非结构化剪枝

- 实验结果 — NVIDIA A100 GPU

Table 1. OPT perplexity results on raw-WikiText2.

OPT - 50%	125M	350M	1.3B
Dense	27.66	22.00	14.62
Magnitude	193.	97.80	1.7e4
AdaPrune	58.66	48.46	32.52
SparseGPT	**36.85**	**31.58**	**17.46**

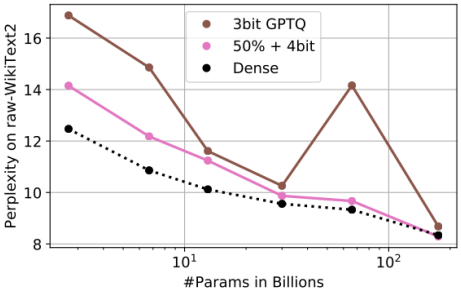
OPT	Sparsity	2.7B	6.7B	13B	30B	66B	175B
Dense	0%	12.47	10.86	10.13	9.56	9.34	8.35
Magnitude	50%	265.	969.	1.2e4	168.	4.2e3	4.3e4
SparseGPT	50%	**13.48**	**11.55**	**11.17**	**9.79**	**9.32**	**8.21**
SparseGPT	4:8	14.98	12.56	11.77	10.30	9.65	8.45
SparseGPT	2:4	17.18	14.20	12.96	10.90	10.09	8.74

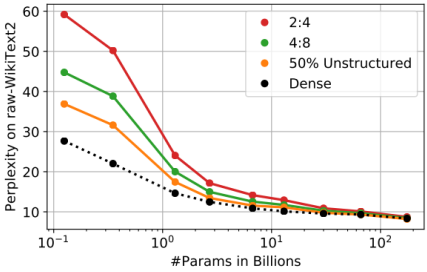
Table 2. ZeroShot results on several datasets for sparsified variants of OPT-175B.

Method	Spars.	Lamb.	PIQA	ARC-e	ARC-c	Story.	**Avg.**
Dense	0%	75.59	81.07	71.04	43.94	79.82	**70.29**
Magnitude	50%	00.02	54.73	28.03	25.60	47.10	**31.10**
SparseGPT	50%	78.47	80.63	70.45	43.94	79.12	**70.52**
SparseGPT	4:8	80.30	79.54	68.85	41.30	78.10	**69.62**
SparseGPT	2:4	80.92	79.54	68.77	39.25	77.08	**69.11**

非结构化剪枝

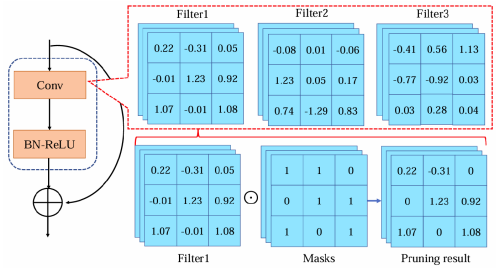
- 推广至半结构化剪枝以及和量化结合





非结构化剪枝

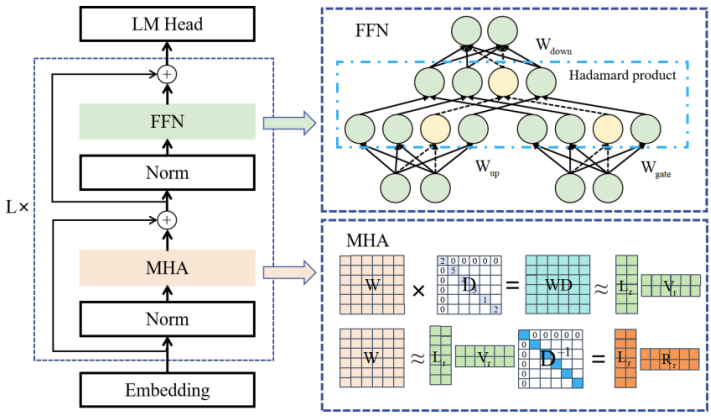
- 网络剪枝与大模型剪枝



目录

结构化剪枝

- LoRAP: Transformer Sub-Layers Deserve Differentiated Structured Compression for Large Language Models, 2024



结构化剪枝

- SlimLLM (Huawei Noah's Ark Lab, China)

SlimLLM: Accurate Structured Pruning for Large Language Models

Jialong Guo [1] **Xinghao Chen** [1] **Yehui Tang** [1] **Yunhe Wang** [1]

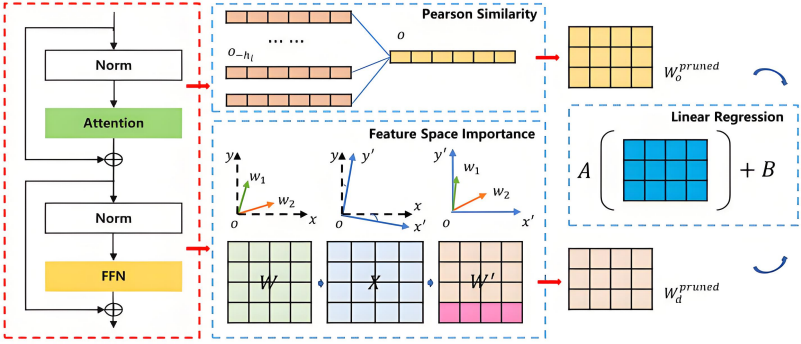
Abstract

Large language models(LLMs) have garnered significant attention and demonstrated impressive capabilities in a wide range of applications. How-

translation to complex reasoning and problem-solving. With the enhancement of model capabilities, there is often a corresponding surge in computational expenses. This often limits the deployment and application of LLMs.

结构化剪枝

- Attention + FFN + Linear Regression



结构化剪枝

- **注意力头剪枝**

$$Q_i = XW_i^Q, K_i = XW_i^k, V_i = XW_i^V$$

$$\text{head}_i = \text{Softmax}\left(\frac{Q_i K_i^\top}{\sqrt{d_k}}\right) V_i$$

$$\text{MHA}(X) = \sum_{i=1}^{h} \text{head}_i W_i^o$$

- $\text{Pearson}(\cdot)$ **相似度**

$$\text{Score}_i = -\text{Pearson}(XW_o, XW_o - X_i W_i^o)$$

结构化剪枝

- 贪婪搜寻算法

Input: pruned attention heads set: S_p={$head_i$, if $head_i$ is pruned.}, unpruned attention heads set: S_{-p}={$head_i$, if $head_i$ is unpruned.}
Output: left attention heads set S_{left}

$S_{left} = S_{-p}$
$O_{-p} = Output(S_{-p})$
$O_{all} = Output(S_p + S_{-p})$
$Sim = Pearson(O_{-p}, O_{all})$
for $h_i \in S_p$ **do**
  for $h_j \in S_{-p}$ **do**
    $O = Output(S_{-p} - h_j + h_i)$
    $s = Pearson(O, O_{all})$
    if $s > Sim$ **then**
      $Sim = s$
      $S_{left} = S_{-p} - h_j + h_i$
    end if
  end for
  $S_{-p} = S_{left}$
end for

结构化剪枝

- FFN 剪枝

$$W' = W_{\text{down}}^{\top} Q$$
$$C_i = \text{sigmoid}(M_i / \bar{M})$$

$$\Downarrow$$

$$I_j^d = \|W_{j1}^{'} C_1, W_{j2}^{'} C_2, ..., W_{jD}^{'} C_D\|_2$$

$$\Downarrow$$

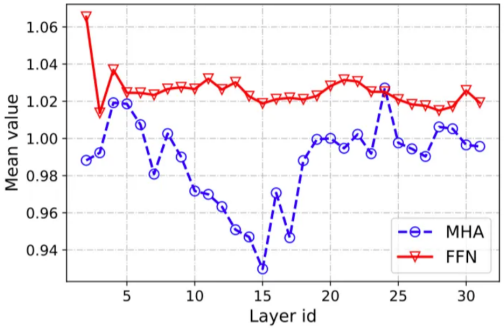
$$I_j = \|X_j\|_2 I_j^d + \|X_{L2} W_{\text{gate}}^j\|_2 + \|X_{L2} W_{\text{up}}^j\|_2$$
$$r_{\text{layer}_i} = r_0 \cdot \text{Softmax}\left(\alpha \cdot \mathbb{E}\left[\frac{X_{i,t} X_{i+1,t}^{\top}}{\|X_{i,t}\|_2 \cdot \|X_{i+1,t}\|_2}\right]\right)$$

结构化剪枝

- 线性回归策略

$$O_i = A_i O_i^{\text{pruned}} + B_i$$

$$O = X_{\text{in}}(A \cdot W_O)^\top + B$$



结构化剪枝

- 零样本实验

Ratio	Method	WikiText2	PTB	BoolQ	PIQA	HellaSwag	WinoGrande	ARC-e	ARC-c	OBQA	Average
Ratio=0%	Llama-7B	12.62	22.14	73.18	78.35	72.99	67.01	67.45	41.38	42.4	63.25
Ratio=20% w/o tune	LLM-pruner	19.09	34.21	57.06	75.68	66.8	59.83	60.94	36.52	40.0	56.69
	LoRAPrune	20.67	34.12	57.98	75.11	65.81	59.90	62.14	34.59	39.98	56.50
	LoRAP	**15.69**	**25.86**	71.93	**76.44**	**69.98**	65.9	60.56	38.48	**40.4**	60.53
	Ours	15.95	26.09	**72.72**	75.95	69.82	**66.06**	**64.48**	**39.33**	40.2	**61.22**
Ratio=20% w/ tune	LLM-pruner	17.58	30.11	64.62	77.2	68.8	63.14	64.31	36.77	39.8	59.23
	LoRAPrune	16.80	28.75	65.62	**79.31**	70.00	62.76	65.87	37.69	39.14	60.05
	LoRAP	16.35	27.06	72.94	76.93	70.9	65.75	64.31	39.93	**41.2**	61.7
	Ours	**15.55**	**26.66**	**74.71**	76.61	**71.23**	**66.54**	**66.96**	**40.61**	40.2	**62.41**
Ratio=50% w/o tune	LLM-pruner	112.44	255.38	52.32	59.63	35.64	53.20	33.50	27.22	33.40	42.13
	LoRAPrune	121.96	260.14	51.78	56.90	36.76	53.80	33.82	26.93	33.10	41.87
	LoRAP	56.96	87.71	57.8	63.82	46.96	57.3	40.36	27.73	36.80	47.25
	Ours	**37.89**	**67.68**	**63.33**	**65.40**	**49.94**	**58.80**	**45.83**	**30.38**	**37.00**	**50.10**
Ratio=50% w/ tune	LLM-pruner	38.12	66.35	60.28	69.31	47.06	53.43	45.96	29.18	35.60	48.69
	LoRAPrune	30.12	50.30	61.88	**71.53**	47.86	55.01	45.13	31.62	34.98	49.71
	LoRAP	30.90	48.84	**63.00**	69.64	54.42	58.41	51.94	32.00	35.80	52.17
	Ours	**26.71**	**42.19**	62.78	68.99	**54.73**	**61.01**	**54.55**	**33.28**	**36.80**	**53.16**

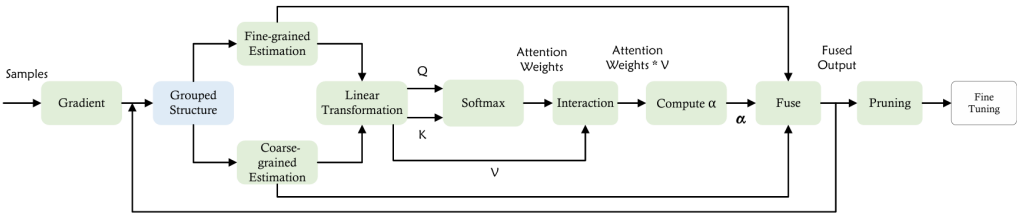
结构化剪枝

- 零样本实验

Ratio	Method	WikiText2	PTB	BoolQ	PIQA	HellaSwag	WinoGrande	ARC-e	ARC-c	OBQA	Average
Ratio=0%	Llama2-7B	12.18	47.25	71.04	78.40	72.96	67.17	69.32	40.53	40.80	62.89
Ratio=20% w/o tune	LoRAP	**15.02**	58.44	69.24	**76.39**	**69.15**	**65.11**	61.99	35.58	38.60	59.44
	Ours	15.70	**56.33**	**69.79**	76.28	68.88	63.54	**65.74**	**39.08**	**39.80**	**60.44**
Ratio=20% w/ tune	LoRAP	**14.67**	57.52	70.89	**78.13**	69.93	**65.67**	65.99	38.48	**39.60**	61.24
	Ours	15.28	**55.46**	**72.29**	78.02	**70.95**	64.88	**67.17**	**38.99**	**39.60**	**61.70**
Ratio=50% w/o tune	LoRAP	60.89	282.22	61.86	62.23	43.98	**55.41**	38.51	27.65	33.00	46.09
	Ours	**38.64**	**141.06**	**62.69**	**64.74**	**45.91**	53.28	**39.73**	**29.01**	**33.2**	**46.93**
Ratio=50% w/ tune	LoRAP	**26.26**	101.22	63.27	**70.78**	**55.14**	**57.85**	52.15	30.97	36.00	**52.31**
	Ours	27.29	**88.28**	**64.19**	69.04	53.60	55.33	**52.53**	**32.08**	**37.40**	52.02

混合粒度剪枝

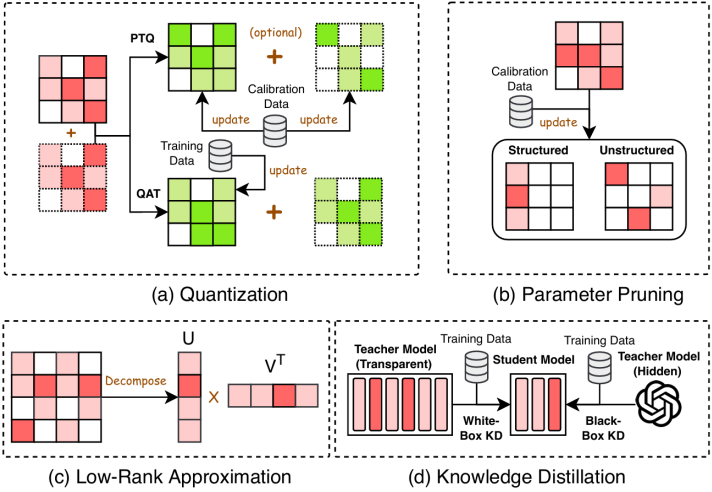
- Toward Adaptive Large Language Models Structured Pruning via Hybrid-grained Weight Importance Assessment, 2025



目录

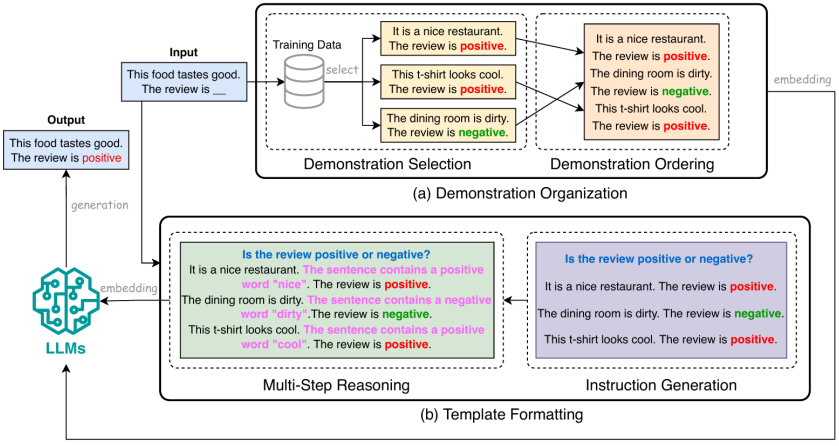
小结

- Efficient Large Language Models: A Survey, 2024



小结

- Efficient Large Language Models: A Survey, 2024



小结

- Optimizing LLMs for Resource-Constrained Environments: A Survey of Model Compression Techniques, 2025
- Efficient Inference for Large Reasoning Models: A Survey, 2025
- Compression Laws for Large Language Models, 2025
- A Survey on Model Compression for Large Language Models, 2024
- Model Compression and Efficient Inference for Large Language Models: A Survey, 2024
- A Survey on Transformer Compression, 2024
- Efficient Large Language Models: A Survey, 2024

Q&A

Thank you!

感谢您的聆听和反馈