浙江大學
ZHEJIANG UNIVERSITY

ICLR

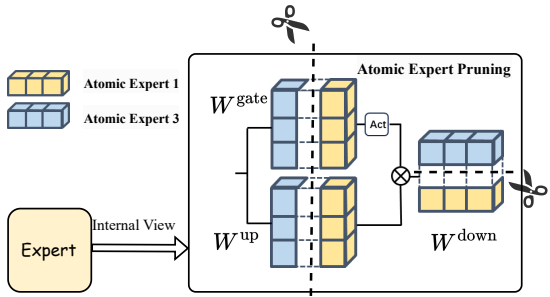
HEAPr: Hessian-based Efficient Atomic Expert Pruning in Output Space

Ke Li[1] Zheng Yang[2] Zhongbin Zhou[3] Feng Xue[4] Zhonglin Jiang[4]
Wenxiao Wang[1*]

[1] School of Software Technology, Zhejiang University
[2] FABU Inc.
[3] Hangzhou Kuaidi Science and Technology Co., Ltd.
[4] Geely Automobile Research Institute (Ningbo) Co., Ltd
{like2248,wenxiaowang}@zju.edu.cn

What is Atomic Expert in MoEs

- **Motivation:** Expert pruning requires **more flexible granularity** to avoid accuracy loss from dropping whole experts.
- **Definition:** An atomic expert is the smallest indivisible unit within an expert..
  - **Construction:** Formed by grouping the j-th hidden dimension across W_{gate}, W_{up}, and W_{down}.



Atomic Expert level

How to Measure Atomic Expert Importance?

- Ideal Framework: Optimal Brain Surgeon (OBS) Approach

$$\Delta\ell = \ell(\theta + \delta\theta) - \ell(\theta) = \nabla\ell(\theta)^\top \delta\theta + \frac{1}{2}\delta\theta^\top H \delta\theta + O(\|\delta\theta\|^3), \tag{1}$$

$$\min_{\delta\theta_q} \ \frac{1}{2}\delta\theta^\top H \delta\theta, \quad \text{s.t. } \delta\theta_q + \theta_q = 0, \tag{2}$$

- The Computational Bottleneck: Expert-level Hessian (H) requires $O((3d_{\text{model}} \cdot d_{\text{inter}})^2)$ space complexity.

Step 1: Atomic Expert Independence

The atomic expert is independent.

Atomic experts are independent, implying the **cross-Hessians** vanish:

$$H_{ij} = \frac{\partial^2 E(\mathbf{x})}{\partial \Theta^{(i)} \partial \Theta^{(j)}} = \frac{\partial^2 \mathbf{e}^{(i)}(\mathbf{x})}{\partial \Theta^{(i)} \partial \Theta^{(j)}} = 0, \quad \forall i \neq j \tag{3}$$

Loss Change The global loss change $\Delta\ell$ decomposes into a sum of local changes:

$$\Delta\ell \approx \frac{1}{2} \sum_{i=1}^{d_{\text{inter}}} (\delta\Theta^{(i)})^T H^{(i)} \delta\Theta^{(i)} \tag{4}$$

Space Complexity Reduction

- Before: $\mathcal{O}\left((3d_{\text{model}} \cdot d_{\text{inter}})^2\right)$
- After: $\mathcal{O}\left((3d_{\text{model}})^2 \cdot d_{\text{inter}}\right)$

Step 2: Shifting to Output Space Optimization

- **Equivalent Form of Pruning Parameters:**

$$\mathbf{e}_{\mathcal{P}}(\mathbf{x}; \Theta_{\mathcal{P}} + \delta\Theta_{\mathcal{P}}) \approx \mathbf{e}_{\mathcal{P}}(\mathbf{x}; \Theta_{\mathcal{P}}) + J_{\mathcal{P}}\delta\Theta_{\mathcal{P}} = \mathbf{0}, \tag{5}$$

- **Optimization:**

$$\min_{\delta\Theta_{\mathcal{P}}} \frac{1}{2} \sum_{i=1}^{d_{\text{inter}}} (\delta\Theta^{(i)})^T H^{(i)} \delta\Theta^{(i)} \quad \text{s.t.} \quad J_{\mathcal{P}}\, \delta\Theta_{\mathcal{P}} + \mathbf{e}_{\mathcal{P}} = \mathbf{0}. \tag{6}$$

- **Fisher Information & Chain Rule:**

$$\mathbb{E}[H] = F = \mathbb{E}\left[(\nabla_{\Theta}\ell)(\nabla_{\Theta}\ell)^T\right] \quad \text{and} \quad \nabla_{\Theta_{\mathcal{P}}}\ell = J_{\mathcal{P}}^{\top}\mathbf{g}_{\mathcal{P}}$$

- **Output Space Importance Metric:**

$$s \approx \mathbb{E}_{x\sim D}\left[\frac{1}{2} e_{\mathcal{P}}^{\top} \mathbb{E}[g_{\mathcal{P}} g_{\mathcal{P}}^{\top}] e_{\mathcal{P}}\right] \tag{7}$$

Step 2: Shifting to Output Space Optimization

Shared Gradient Covariance (within one expert).

All atomic experts within the same expert share the identical output gradient:

$$\frac{\partial \ell}{\partial \mathbf{e}^{(i)}(\mathbf{x})} = \frac{\partial \ell}{\partial E(\mathbf{x})}, \quad \forall i \in \{1, \ldots, d_{\text{inter}}\},\ \mathbf{e}^{(i)} \in E. \tag{8}$$

Global Ranking Rank globally by $s = \mathbb{E}_{\mathbf{x}\sim D}\left[\frac{1}{2}\,\mathbf{e}_{\mathcal{P}}^{\top}\,\mathbb{E}[\mathbf{g}_{\mathcal{P}}\mathbf{g}_{\mathcal{P}}^{\top}]\,\mathbf{e}_{\mathcal{P}}\right]$ and prune bottom $r\%$

Space Complexity Reduction

- Before: $\mathcal{O}\big((3d_{\text{model}} \cdot d_{\text{inter}})^2\big)$
- After Step 1: $\mathcal{O}\left((3d_{\text{model}})^2 \cdot d_{\text{inter}}\right)$
- After Step 2: $\mathcal{O}(d_{\text{model}}^2)$

HEAPr Algorithm

Algorithm 1 HEAPr: Hessian-based Efficient Atomic Expert Pruning in Output Space

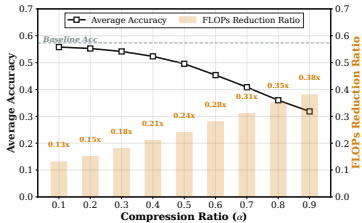
Require: MoE model f_θ, calibration set $\mathcal{D}$, pruning ratio r
Ensure: Pruned model $f_{\theta'}$

for each expert E_i **do** ▷ Stage 1: Gradient Covariance Estimation
    Collect routed tokens $\mathcal{T}_i$
    Compute shared gradient $\mathbf{g}_{E_i} = \frac{\partial \ell}{\partial E_i}$
    Compute $\bar{\boldsymbol{G}}_i = \frac{1}{|\mathcal{T}_i|} \sum_{\mathbf{x} \in \mathcal{T}_i} \mathbf{g}_{E_i}(\mathbf{x}) \mathbf{g}_{E_i}(\mathbf{x})^\top$ ▷ Space complexity $\mathcal{O}(d^2)$
end for
for each atomic expert $\mathbf{e}_k$ in E_i **do** ▷ Stage 2: Importance Computation
    Compute $\bar{s}_k = \frac{1}{|\mathcal{T}_i|} \sum_{\mathbf{x} \in \mathcal{T}_i} \frac{1}{2} \mathbf{e}_k(\mathbf{x})^\top \bar{\boldsymbol{G}}_i \mathbf{e}_k(\mathbf{x})$
end for
Global rank $\{\bar{s}_k\}$ and prune lowest $r\%$ across all experts
return Pruned model $f_{\theta'}$

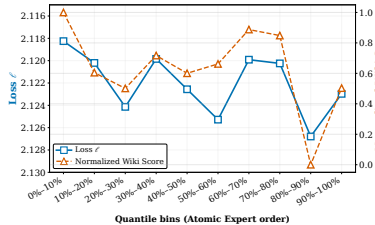
Performance

Ratio	Method	Wiki↓	PTB↓	Openb.	ARC_e	WinoG.	HellaS.	ARC_c	PIQA	MathQA	Avg.↑
DeepSeekMoE-16B-Base											
0%	Original	6.38	9.47	0.32	0.76	0.71	0.58	0.45	0.79	0.32	0.56
20%	NAEE	9.44	15.02	**0.32**	0.71	0.66	0.55	0.40	0.77	0.29	0.53
	MoE-I^2	7.69	11.59	0.26	0.71	0.68	0.49	0.38	0.73	0.29	0.50
	MoE-SVD	6.92	10.48	0.31	0.75	0.70	0.53	0.42	0.76	0.31	0.54
	D^2-MoE	6.84	11.10	0.30	0.74	0.69	0.55	0.41	0.76	0.31	0.54
	HEAPr	**6.54**	**9.88**	**0.32**	**0.76**	**0.71**	**0.57**	**0.45**	**0.79**	**0.32**	**0.56**
40%	NAEE	8.55	14.47	0.23	0.67	0.67	0.41	0.32	0.69	0.26	0.46
	MoE-I^2	9.73	15.75	0.23	0.64	0.66	0.41	0.31	0.68	0.26	0.45
	D^2-MoE	7.93	14.07	0.26	0.69	0.65	0.45	0.36	0.72	0.28	0.49
	HEAPr	**6.80**	**10.86**	**0.30**	**0.74**	**0.69**	**0.52**	**0.41**	**0.76**	**0.30**	**0.53**
Qwen1.5-MoE-A2.7B-Chat											
0%	Original	8.12	12.97	0.31	0.70	0.66	0.59	0.40	0.79	0.35	0.54
25%	MC-SMoE	12.76	17.45	0.25	0.65	0.65	0.53	0.37	-	-	-
	HC-SMoE	11.62	16.39	0.27	0.66	0.63	0.55	0.35	**0.76***	0.29*	0.50
	Sub-MoE	9.48	14.84	0.30	**0.69**	0.66	**0.56**	0.37	-	-	-
	HEAPr	**8.31**	**14.12**	**0.32**	**0.69**	**0.67**	**0.56**	**0.38**	**0.76**	**0.35**	**0.53**
50%	MC-SMoE	5e2	1e3	0.18	0.33	0.52	0.29	0.19	-	-	-
	HC-SMoE	25.50	38.18	0.23	0.61	**0.65**	**0.47**	**0.35**	0.58*	0.23*	0.45
	Sub-MoE	17.51	29.00	0.25	0.58	0.58	0.46	0.25	-	-	-
	HEAPr	**9.24**	**17.58**	**0.27**	**0.64**	0.64	0.46	0.33	**0.71**	**0.33**	**0.48**
Qwen3-30B-A3B											
0%	Original	8.64	15.40	0.34	0.79	0.71	0.60	0.54	0.79	0.59	0.62
25%	HC-SMoE	18.86	31.11	0.22	0.64	0.61	0.40	0.35	0.59*	0.41*	0.46
	Sub-MoE	13.59	23.48	0.25	0.70	0.66	0.47	0.44	-	-	-
	HEAPr	**9.10**	**16.80**	**0.33**	**0.76**	**0.70**	**0.55**	**0.49**	**0.78**	**0.50**	**0.59**
50%	HC-SMoE	72.33	162.99	0.13	0.44	0.50	0.29	0.23	0.44*	0.32*	0.34
	Sub-MoE	21.05	43.19	0.23	**0.68**	0.63	**0.41**	0.40	-	-	-
	HEAPr	**11.22**	**26.29**	**0.25**	0.67	**0.63**	0.38	**0.41**	**0.67**	**0.36**	**0.48**
Qwen2-57B-A14B											
0%	Original	5.12	9.18	0.33	0.75	0.74	0.63	0.46	0.81	0.39	0.59
40%	NAEE	6.81	11.34	0.31	0.73	0.73	0.55	0.46	0.76	0.36	0.55
	MoE-I^2	24.90	77.05	0.26	0.70	0.46	0.71	0.41	0.75	0.30	0.51
	D^2-MoE	8.19	11.23	**0.33**	**0.75**	**0.75**	0.61	0.45	0.79	0.36	0.58
	HEAPr	**5.92**	**9.34**	**0.33**	**0.75**	0.74	**0.63**	**0.46**	**0.81**	**0.39**	**0.59**

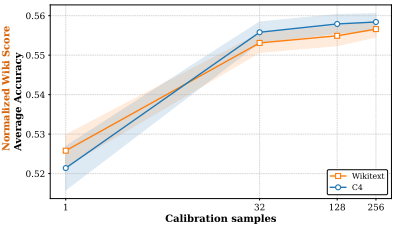
More Result



(a) Accuracy–FLOPs trade-off vs. pruning ratio



(b) Consistency of importance s with actual loss increase $\Delta\ell$



(c) Calibration data & sample-size sensitivity

Future Work

- Hessian-based Parameter Compensation: Beyond atomic expert pruning, future will focus on implementing efficient, training-free parameter updates within the OBS framework.

Thank you