Bypass Back-propagation: Optimization-based Structural Pruning for Large Language Models via Policy Gradient

ACL 2025 VIENNA JULY 27 - AUGUST 1

Yuan Gao*, Zujing Liu*, Weizhong Zhang*, Bo Du, Gui-Song Xia†

Wuhan University, Fudan University

Ethan.Y.Gao@gmail.com

https://github.com/ethanygao/backprop-free_LLM_pruning

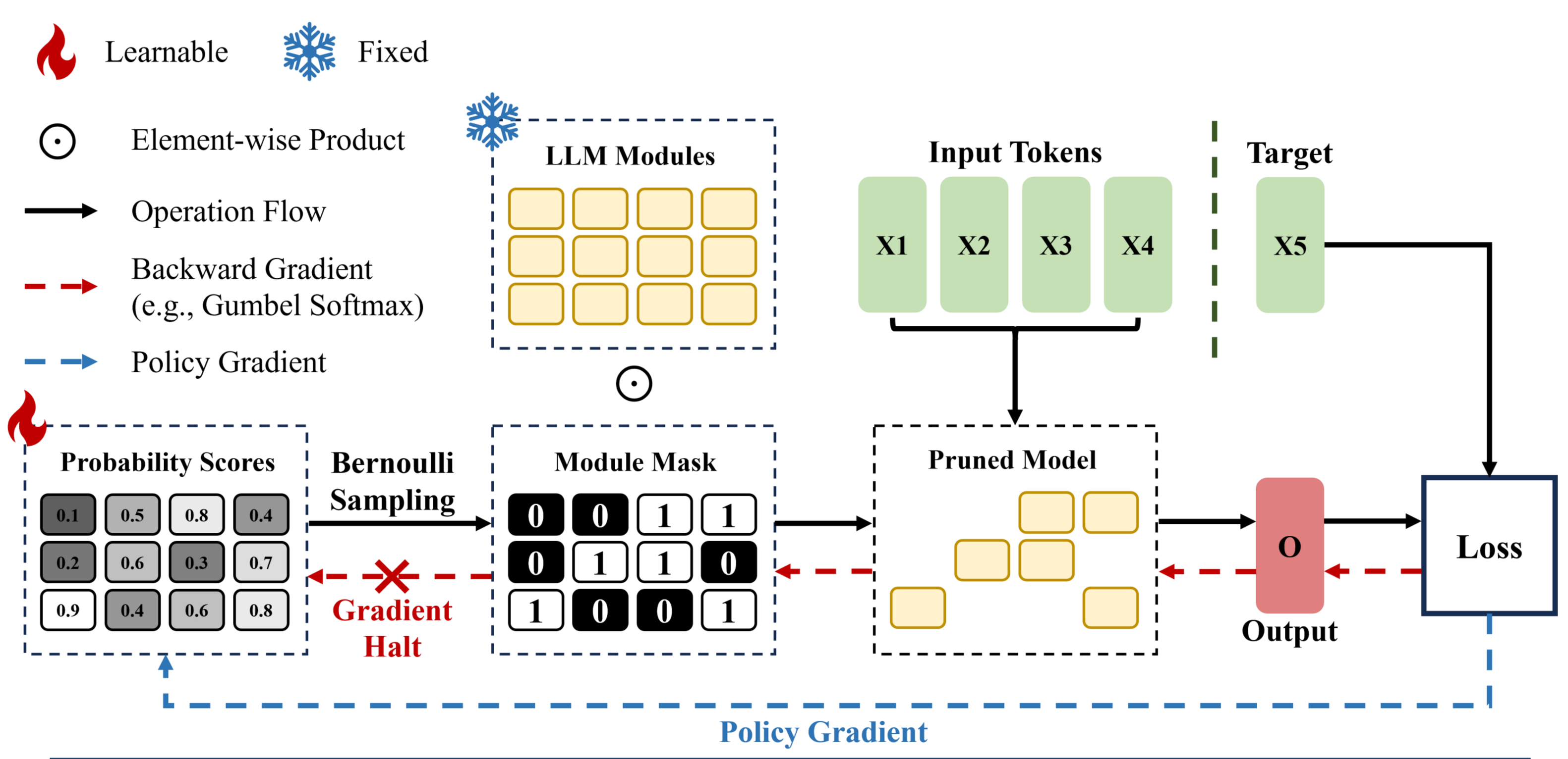
Introduction

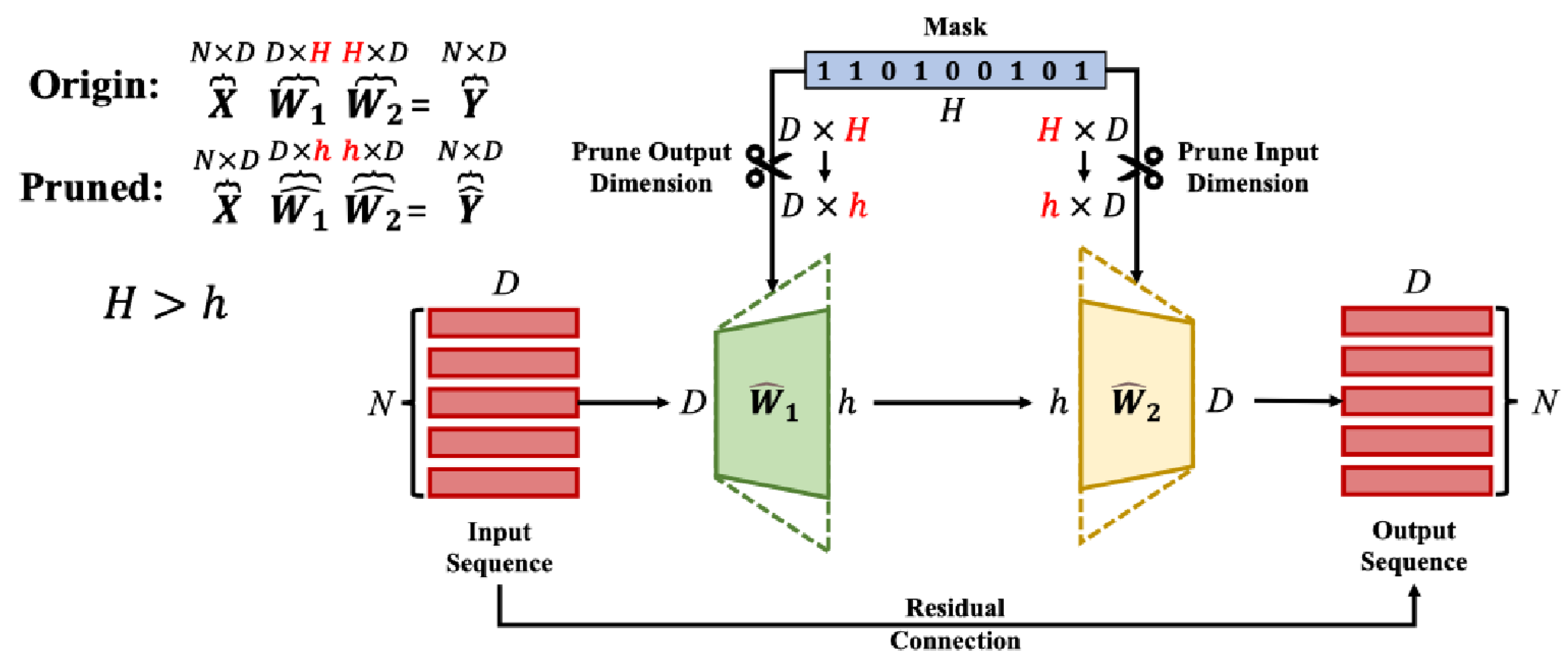
Large Language Models (LLMs) face efficiency challenges for deployment due to their massive parameters.

- Existing LLM metric-based pruning **relies on heuristic metrics**, which is **computationally inexpensive**, but often leads to **suboptimal performance**, especially at high pruning rates.
- Pre-LLM optimization-based methods **require backpropagation**, which can lead to **better performance**, but is **computationally expensive** for LLMs.

Can we attain the **performance of optimization-based methods** while preserving a **similar inexpensive resources with the metrics-based** methods?

Method





1. Pruning is a Binary-Optimization Problem

$$\mathbf{m}=\{\mathbf{m}_i\}_{i=1}^n \in \{0,1\}^n \qquad \mathbf{w}=\{\mathbf{w}_i\}_{i=1}^n$$

$$\min_{\mathbf{m}} \mathcal{L}(\mathcal{D};\mathbf{w}\odot\mathbf{m}) := \frac{1}{N}\sum_{i=1}^{N}\ell(f(\mathbf{x}_i;\mathbf{w}\odot\mathbf{m}),\mathbf{y}_i),$$

$$\text{s.t. } \|\mathbf{m}\|_1 \le rn \text{ and } \mathbf{m}\in\{0,1\}^n.$$

2. Pruning via Probabilistic Mask Modeling

$$\mathbf{s}=\{\mathbf{s}_i\}_{i=1}^n \in [0,1]^n \qquad p(\mathbf{m}|\mathbf{s}) = \prod_{i=1}^{n}(s_i)^{m_i}(1-s_i)^{1-m_i}.$$

$$\min_{\mathbf{s}} \mathbb{E}_{p(\mathbf{m}|\mathbf{s})}\mathcal{L}(\mathcal{D};\mathbf{w}\odot\mathbf{m}),$$

$$\text{s.t. } \mathbf{1}^\top\mathbf{s}\le rn \text{ and } \mathbf{s}\in[0,1]^n.$$

Optimization Object in Expectation

$$\Phi(\mathbf{s}) = \mathbb{E}_{p(\mathbf{m}|\mathbf{s})}\mathcal{L}(\mathbf{m}) = \int p(\mathbf{m}|\mathbf{s})\mathcal{L}(\mathbf{m})\mathrm{d}\mathbf{m},$$

$$\text{s.t. } \mathbf{1}^\top\mathbf{s}\le rn \text{ and } \mathbf{s}\in[0,1]^n.$$

3. Policy Gradient Optimization

$$\nabla_{\mathbf{s}}\Phi(\mathbf{s}) = \int \mathcal{L}(\mathbf{m})\nabla_{\mathbf{s}}p(\mathbf{m}|\mathbf{s}) + \underbrace{p(\mathbf{m}|\mathbf{s})\nabla_{\mathbf{s}}\mathcal{L}(\mathbf{m})}_{=0}\,\mathrm{d}\mathbf{m}$$

$$= \int \mathcal{L}(\mathbf{m})p(\mathbf{m}|\mathbf{s})\nabla_{\mathbf{s}}\log(p(\mathbf{m}|\mathbf{s}))\mathrm{d}\mathbf{m}$$

$$= \mathbb{E}_{p(\mathbf{m}|\mathbf{s})}\mathcal{L}(\mathbf{m})\nabla_{\mathbf{s}}\log(p(\mathbf{m}|\mathbf{s})).$$

Stochastic Gradient Descent Algorithm

$$\mathbf{s} \leftarrow \mathbf{proj}_{\mathcal{C}}(\mathbf{z}),$$

$$\mathbf{z} := \mathbf{s} - \eta\mathcal{L}(\mathcal{D}_B;\mathbf{w}\odot\mathbf{m})\nabla_{\mathbf{s}}\log(p(\mathbf{m}|\mathbf{s})).$$

Variance Reduction via Moving Average Baseline

$$\mathbf{s} \leftarrow \mathbf{proj}_{\mathcal{C}}(\mathbf{z}) \text{ with } \mathbf{z} := \mathbf{s} - \eta\left[\frac{1}{N_s}\sum_{i=1}^{N_s}\left(\mathcal{L}(\mathcal{D}_B;\mathbf{w}\odot\mathbf{m}^{(i)})-\delta\right)\nabla_{\mathbf{s}}\log(p(\mathbf{m}^{(i)}|\mathbf{s}))\right].$$

$$\delta \leftarrow \frac{T-1}{T}\delta + \frac{1}{N_sT}\sum_{i=1}^{N_s}\mathcal{L}(\mathcal{D}_B;\mathbf{w}\odot\mathbf{m}^{(i)}).$$

Experiments

Method	PruneRate	LLaMA		LLaMA-2		LLaMA-3	Vicuna	
		7B	13B	7B	13B	8B	7B	13B
Dense	0%	12.62	10.81	12.19	10.98	14.14	16.24	13.50
LLM-Pruner	30%	38.41	24.56	38.94	25.54	40.18	48.46	31.29
SliceGPT		-	-	40.40	30.38	183.94	52.23	57.75
Bonsai		30.49	26.24	39.01	24.23	80.89	44.28	54.16
Wanda-sp		98.24	25.62	49.13	41.57	92.14	57.60	80.74
Ours		**25.61**	**19.70**	**28.18**	**21.99**	**38.99**	**34.51**	**26.42**
LLM-Pruner	40%	72.61	36.22	68.48	37.89	70.60	88.96	46.88
SliceGPT		-	-	73.76	52.31	353.09	89.79	130.86
Bonsai		60.65	58.17	69.18	50.97	204.61	95.32	272.10
Wanda-sp		110.10	165.43	78.45	162.50	213.47	85.51	264.22
Ours		**42.96**	**28.12**	**39.81**	**31.52**	**63.85**	**51.86**	**43.59**
LLM-Pruner	50%	147.83	67.94	190.56	72.89	145.66	195.85	91.07
SliceGPT		-	-	136.33	87.27	841.20	160.04	279.33
Bonsai		275.63	148.92	216.85	146.38	440.86	180.75	424.33
Wanda-sp		446.91	406.60	206.94	183.75	413.86	242.41	373.95
Ours		**72.02**	**49.08**	**65.21**	**52.23**	**119.75**	**71.18**	**68.13**

Table1. Results (perplexity) on channels and heads pruning. Our method is initialized by Wanda-sp. All the methods are calibrated using the C4 dataset and validated on the WikiText2 dataset w.r.t. perplexity.

Method	PruneRate	PPL ↓	PIQA	HellaSwag	WinoGrande	ARC-e	ARC-c	Average
Dense	0%	14.14	79.71	60.19	72.61	80.09	50.34	68.59
LLM-Pruner	30%	40.18	71.38	37.84	55.64	57.78	27.21	49.97
SliceGPT		183.94	68.34	**53.92**	57.22	49.41	28.07	51.39
Bonsai		80.89	64.53	36.10	55.09	47.64	22.52	45.18
Wanda-sp		92.14	59.74	31.46	52.64	44.02	19.88	41.55
Ours		**38.99**	**72.25**	43.56	**59.04**	**59.85**	**29.44**	**52.83**
LLM-Pruner	40%	70.60	66.26	31.90	54.06	49.74	22.52	44.90
SliceGPT		353.09	61.53	**39.98**	52.80	36.66	**25.17**	43.23
Bonsai		204.61	58.81	29.43	48.93	33.21	18.15	37.71
Wanda-sp		213.47	56.58	27.46	50.35	32.07	17.06	36.70
Ours		**63.85**	**67.63**	37.36	**56.91**	**50.67**	24.91	**47.50**
LLM-Pruner	50%	145.65	61.15	29.10	**51.93**	39.98	19.36	40.30
SliceGPT		841.20	56.37	**32.66**	48.38	32.45	**22.10**	38.39
Bonsai		440.86	55.66	26.94	50.51	30.64	17.83	36.32
Wanda-sp		413.86	55.39	27.07	49.72	29.59	18.26	36.01
Ours		**119.75**	**62.51**	30.89	51.85	**41.12**	20.65	**41.40**

Table2. Perplexity (PPL) and zero-shot accuracies (%) of LLaMA-3-8B for 5 zero-shot tasks.

Method	PruneRate	Perplexity	PruneRate	Perplexity	PruneRate	Perplexity
LLM-Pruner	30%	38.94	40%	68.48	50%	190.56
SliceGPT		40.40		73.76		136.33
Bonsai		39.01		69.18		216.85
Wanda-sp		49.13		78.45		206.94
Ours (Random Init)	30%	37.24	40%	60.16	50%	160.75
Ours (Random-Prog. Init)		<u>31.43</u>		<u>49.86</u>		<u>86.55</u>
Ours (LLM-Pruner Init)	30%	35.75	40%	65.32	50%	116.80
Ours (Wanda-sp Init)		**28.18**		**39.81**		**65.21**

Table 3: Channels and heads pruning results with different initializations on LLaMA-2-7B. Bold and Underscored denote the first and second best results, respectively.

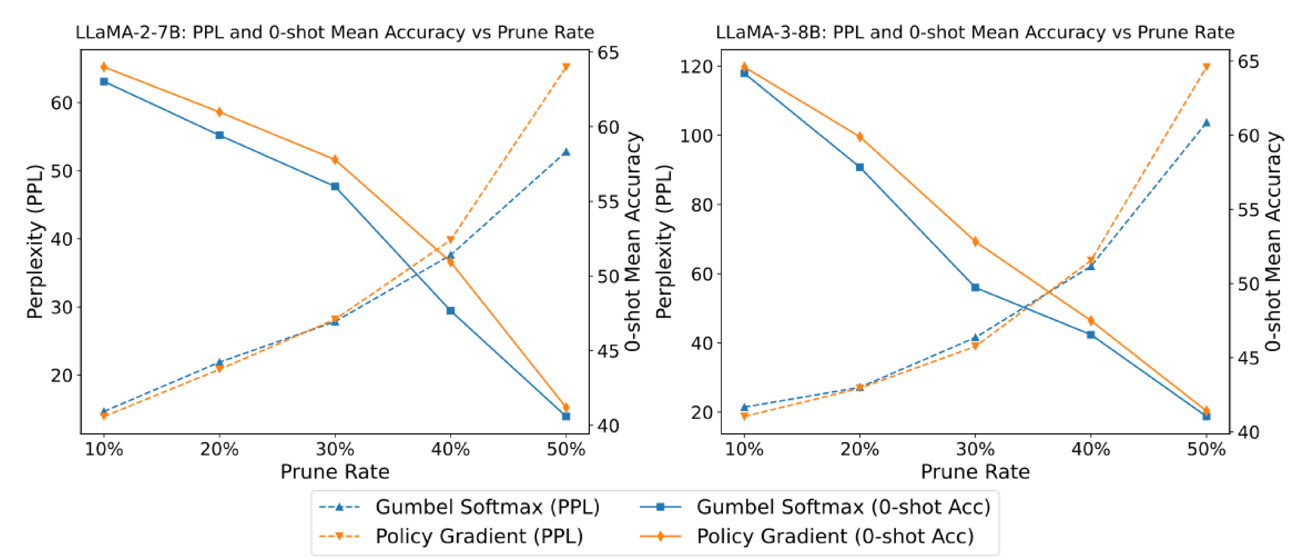


Figure 3: Comparison of Policy Gradient and Gumbel Softmax.

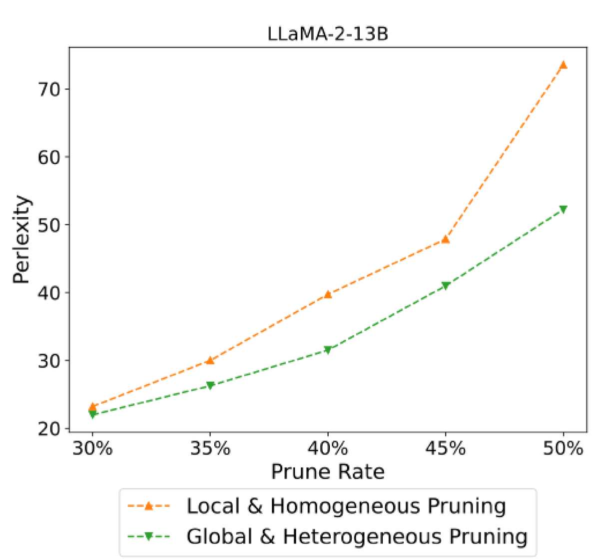


Figure 4: Global vs. local pruning.