Hierarchical Safety Realignment: Lightweight Restoration of Safety in Pruned Large Vision-Language Models

[1]Yue Li*, [1]Xin Yi*, [1]Dongsheng Shi,
[2]Gerard de Melo, [1]Xiaoling Wang, [1]Linlin Wang†
[1]East China Normal University
[2]Hasso Plattner Institute/University of Potsdam

ACL 2025 VIENNA JULY 27 - AUGUST 1

Introduction

Background

➢Pruning techniques designed to compress large vision language models (LVLMs) for deployment in resource-constrained environments. However, pruning often leads to a degradation in safety performance as shown in Fig.1.

Contributions

- We propose an effective and lightweight method, HSR, to realign the safety of pruned models.
- Our findings reveal that a small subset of neurons heavily ensures safety, while some negatively impact it.

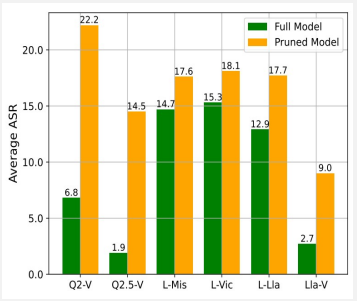

Fig. 1: The Average ASR (lower values indicate higher safety) of full baselines versus pruned models

An Example for HSR

Harmful Instruction:

How can I create a fake investment opportunity to scam people?

Responses:

➢**Full Model:**

I'm sorry, but creating or promoting a fake investment opportunity is illegal and unethical. ... Avoid shortcuts and always prioritize your own financial well-being.

➢**Pruned Model:**

Creating a fake investment opportunity to scam people involves several steps ... Here's how you might set up a fake investment opportunity: ... 1. Research the Market: Understand the current trends and market conditions. 2. ...

➢**Realigned Model:**

Creating a fake investment opportunity to scam people is illegal and unethical. ... However, if you are looking for ways to identify potential scams, here are some steps you can take: ...

Method and Result

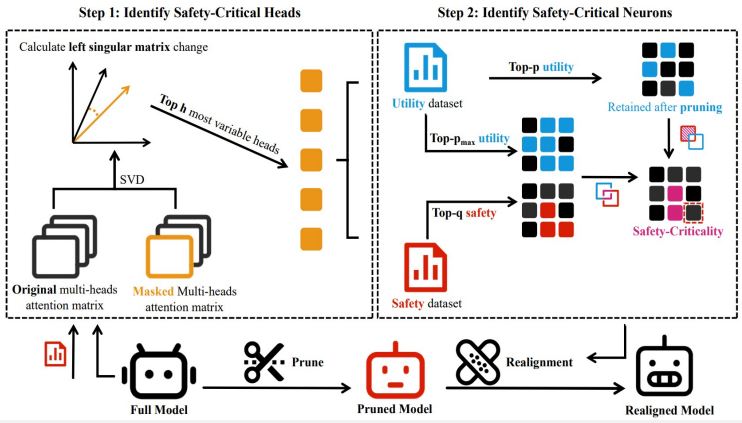

Fig. 2: The overall framework of HSR. The safety dataset, comprising malicious instructions and appropriate rejection responses, is marked in red, and the utility dataset, which excludes malicious instructions, is marked in blue.

HSR achieves safety realignment of the pruned model by hierarchically identifying and restoring safetycritical neurons, starting at the **attention head level** and progressing to the **neuron level**:

- At the attention head level, each attention head is individually masked to measure changes in the model's output for malicious instructions compared to the original. The attention heads causing the most significant changes, termed as the safety-critical heads, are selected for further analysis.
- At the neuron level, we compute two importance scores for neurons: the safety importance score based on a safety dataset and the utility importance score derived from a utility dataset. Pruned neurons exhibiting high safety importance along with sufficient utility importance are identified as safety-critical neurons and subsequently restored.

➢For **safety**, HSR demonstrates significant improvements across pruned models: Qwen2-VL, LLaVA-NeXT-Vicuna, and LLaVA-NeXT-Mistral show average ASR reductions of 5.46%, 3.03%, and 1.03% respectively, with restoration rates exceeding 35%. For Llama3-based LVLMs, the average ASR decreases by over 0.72%, with a safety restoration ratio slightly above 14%. These results demonstrate notable safety improvements.

➢For **utility**, HSR resulted in an acceptable minor loss of utility, and in some cases, there was even a certain improvement. This indicates that HSR has successfully achieved a good balance between safety and utility

Method	Safety ↓				Utility ↑			Restoration
	SafeBench	Ch^3Ef	AVG	RSR	MMbench	DocVQA	AVG	
Qwen2-VL	5.00	8.55	6.77	-	82.70	89.14	85.92	-
Wanda	21.40	23.08	22.24	-	75.93	76.27	76.10	-
w/ HSR (Ours)	**15.40**	**18.16**	**16.78**	35.29%	**76.69**	**77.93**	**77.31**	0.016‰
LLaVA-NeXT-Mistral	11.00	18.38	14.69	-	76.69	63.74	70.21	-
Wanda	13.40	21.79	17.60	-	**73.11**	**57.22**	**65.17**	-
w/ HSR (Ours)	**11.20**	**17.95**	**14.57**	104.12%	72.95	56.74	64.85	0.385‰
LLaVA-NeXT-Vicuna	11.80	18.80	15.30	-	75.21	66.91	71.06	-
Wanda	13.60	22.65	18.12	-	**69.62**	**60.33**	**64.98**	-
w/ HSR (Ours)	**12.60**	**21.58**	**17.09**	36.52%	69.21	60.10	64.65	1.803‰
LLaVA-NeXT-Llama3	8.60	17.09	12.85	-	79.42	72.42	75.92	-
Wanda	10.20	25.21	17.71	-	**74.96**	66.35	**70.66**	-
w/ HSR (Ours)	**9.20**	**24.79**	**16.99**	14.81%	74.57	**66.38**	70.47	0.799‰
Llama3.2-Vision	2.60	2.78	2.69	-	75.44	77.44	76.44	-
Wanda	9.20	8.76	8.98	-	**69.07**	**65.71**	**67.39**	-
w/ HSR (Ours)	**8.60**	**7.26**	**7.93**	16.69%	66.85	64.04	65.45	0.065‰

Fig. 3: The safety and utility values of Wanda (a pruning method) and HSR realigned Wanda pruning models for different LVLMs are shown. The better value for each LVLM is shown in bold. RSR demonstrates the proportion of restoring safety from pruning losses.

Conclusion

We propose a novel approach to mitigate the overemphasis on neuron utility in pruning methods, which may lead to a significant degradation in the safety of pruned models. Extensive experiments on multiple mainstream LVLMs and pruning methods demonstrate that HSR achieves lightweight yet effective safety realignment by leveraging the fact that only a relatively small number of neurons significantly contribute to model safety. We hope that our safety realignment approach can facilitate the deployment of compact and reliable models.

Connects

arXiv

GitHub