MMNeuron: Discovering Neuron-Level Domain-Specific Interpretation in Multimodal Large Language Model

Jiahao Huo, Yibo Yan, Boren Hu, Yutao Yue, Xuming Hu

The Hong Kong University of Science and Technology (Guangzhou)
The Hong Kong University of Science and Technology, Tongji University

❖ Motivation & Key Result

➢ Motivation

- There exist language-specific neurons in multilingual LLMs, whose activation or deactivation determines the output language of models.
- There are obvious gaps between different image domains (Auto Driving, Remote Sensing, Document, etc.), and maybe some neurons are responsible for processing images from specific domain.

➢ Key Result

- We analyze the impact of domain-specific neurons, indicating that both LLaVA-NeXT and InstructBLIP do not fully utilize domain-specific information in particular domains.
- We compare features from various domains through the lens of domain-specific neurons, revealing that images from different domains vary in conceptual depth.
- We propose a three-stage framework of language models in MLLMs when processing projected image features, shedding light on the internal mechanisms by which image features align with word embeddings.

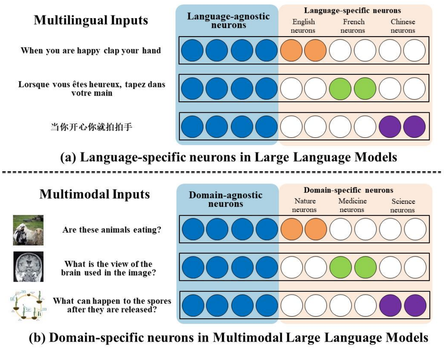
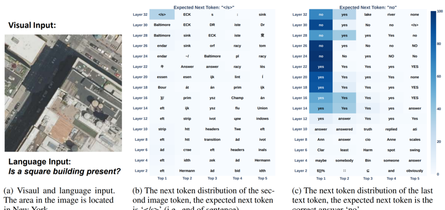


Figure 1. Illustration of domain-specific neurons.



(a) Visaul and language input. The area in the image is located in New York.

(b) The next token distribution of the second image token, the expected next token is '</s>' (i.e., end of sentence).

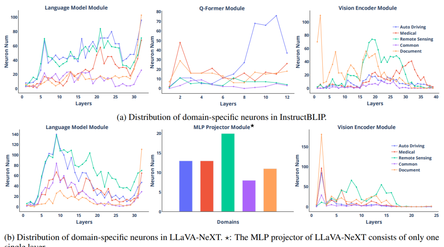
(c) The next token distribution of the last text token, the expected next token is the correct answer 'no'.

Figure 2. Case study of our VQA task.

❖ Main Experimental Results

➢ Distribution of domain-specific neurons

- Two obvious turning points can be observed in language model.
- Domain-specific neurons are mainly distributed in shallow and intermediate layers within vision encoders.



(a) Distribution of domain-specific neurons in InstructBLIP.

(b) Distribution of domain-specific neurons in LLaVA-NeXT. *: The MLP projector of LLaVA-NeXT consists of only one single layer.

Figure 3. Layer-wise Distribution of domain-specific neurons.

➢ Perturbation for Performance in VQA

- Manipulating domain-specific neurons properly will result in a 10% change of accuracy at most
- In some cases, removing domain-specific information seems to benefit the target task.

Model	Deactivated Module(s)	VQAv2	PMC-VQA	LingoQA	RS-VQA	DocVQA
LLaVA-NeXT	None	74.9	34.4	20.6	42.5	59.2
	Vision Encoder	75.8	**34.3**	24.6	42.1	58.3
	MLP Projector	74.9	34.4	24.2	42.5	59.2
	LLM	75.7	34.5	24.2	41.0	59.0
	All	**73.5**	34.5	**24.2**	**38.5**	**57.0**
InstructBLIP	None	66.1	28.1	20.6	34.7	24.0
	Vision Encoder	**66.9**	31.0	21.8	34.8	23.8
	Q-Former	67.1	32.4	20.0	**33.1**	24.6
	LLM	67.1	32.6	24.2	35.5	24.4
	All	68.6	**30.9**	**18.0**	33.6	**23.8**

Table 1. Accuracy (%) of LLaVA-NeXT and InstructBLIP on selected domains with corresponding domain specific neurons deactivated.

❖ MMNeuron Framework

➢ Overall Framework

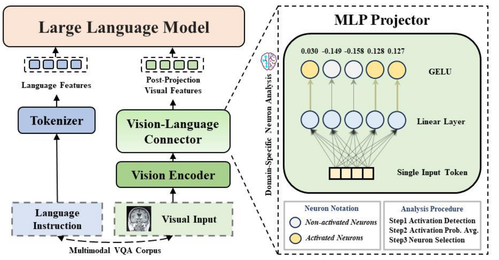


Figure 4. Overall framework of our proposed method.

The activation probability of a neuron u in domain D_l is:

$$p_{u,l} = \frac{M_{u,l}}{N_{u,l}}, \tag{1}$$

Then the probability distribution (normalized through L1 normalization) of neuron u across all domains is:

$$P'_u = (p'_{u,1}, p'_{u,2}, ..., p'_{u,k}), \quad \text{where} \quad p'_{u,i} = \frac{p_{u,i}}{\sum_{j=1}^{k} p_{u,j}}. \tag{2}$$

Finally, calculate its corresponding entropy as domain activation probability entropy (DAPE) as:

$$DAPE_u = -\sum_{j=1}^{k} p_{u,j} \log p_{u,j}. \tag{3}$$

Figure 5. Calculation of Domain Activation Probability Entropy(DAPE), we select top 1% neurons with the smallest DAPE.

❖ Further Analysis

➢ Perturbation for Hidden States

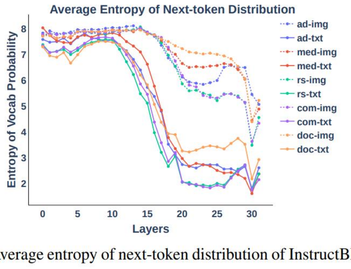
- Deactivating domain-specific neurons causes a large perturbation to LLaVA-NeXT and Instruct-BLIP's hidden states.
- Both LLaVA and InstructBLIP fail to take full advantage of the domain-specific information in specific domains, which may limit their cross-domain ability.

Baseline	Module	VQAv2	PMC-VQA	LingoQA	DocVQA	RS-VQA
LLaVA-NeXT	Random (Avg.)	8.41	18.90	16.04	21.81	32.76
	LLM	0.01	0.01	0.02	0.10	0.02
	Vision Encoder	17.19	30.98	35.74	46.75	49.90
	MLP Projector	0.0	0.0	0.0	0.0	0.0
	All	**17.19**	**30.98**	**35.74**	**46.75**	**49.90**
InstructBLIP	Random (Avg.)	5.13	8.15	8.57	14.85	9.91
	LLM	6.84	12.13	9.62	7.80	11.98
	Vision Encoder	2.44	17.93	5.33	26.11	23.76
	Q-Former	2.93	11.61	6.95	14.58	6.52
	All	**8.00**	**24.84**	**12.77**	**29.04**	**26.58**

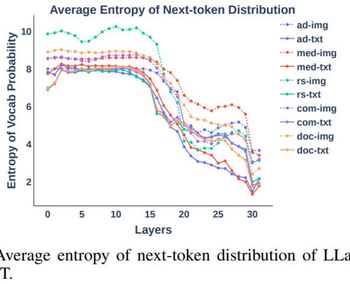
Table 2. The deviation (%) of hidden states of MLLMs' last layer after deactivating domain-specific neurons.

➢ Three-stage mechanism of LLM understanding multimodal features.

- 1) In the first several layers, projected features are further aligned with word space. Around the turning point, the multimodal features are embedded into a uniform representation space.
- 2) Transitioning into the second phase, features are further generalized and understood by language models, where domain-specific neurons decrease sharply.
- 3)In the third stage, language models generate responses to the input, showing a rise of neurons specific to target tasks.



(a) Average entropy of next-token distribution of InstructBLIP.



(b) Average entropy of next-token distribution of LLaVA-NeXT.

Figure 5. The average entropy of next token probability distribution for image and text tokens.