NPM3D Project Start-Up Report Zhe HUANG(29/02/2024)
LiDAR-LLM: Exploring the Potential of LLM for 3D LiDAR Understanding

1 Summary of The Article

The article discusses LiDAR-LLM, a language model innovatively developed for interpreting 3D scenes, notably using sparse outdoor LiDAR data. LiDAR technology, which measures distances with laser light, has been challenging for existing models to grasp fully. This paper's breakthrough is in framing 3D scene understanding as a language modeling problem, encompassing 3D captioning, grounding, and Q&A tasks.

Addressing the lack of LiDAR-text data, a novel three-stage training strategy is proposed to align 3D modalities with language embeddings. A key innovation is the View-Aware Transformer (VAT) that links the 3D encoder and language model, enhancing spatial orientation and visual feature comprehension.

Experimental results validate LiDAR-LLM's adeptness in interpreting instructions for 3D scenes and in complex spatial reasoning, scoring notably in BLEU-1 for 3D captioning, classification accuracy, and BEV mIoU for grounding tasks. These metrics suggest LiDAR-LLM's robust performance in understanding and reasoning about 3D spatial configurations.

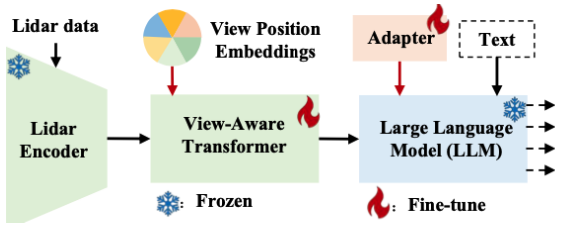
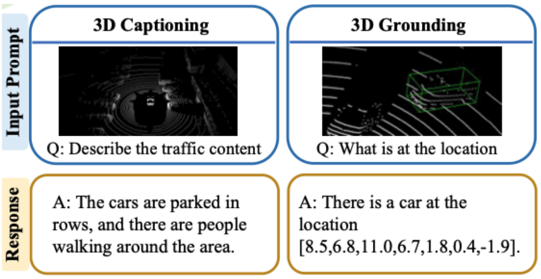


Figure 1: Characteristics of LiDAR-LLM



Q: Describe the traffic content

Q: What is at the location

A: The cars are parked in rows, and there are people walking around the area.

A: There is a car at the location [8.5,6.8,11.0,6.7,1.8,0.4,-1.9].

Figure 2: A little Example

2 Initial Implementation Tests

The article mentioned that they used the standard pre-trained 3D detector, CenterPoint-Voxel following its default settings. I tried to load OpenPCDet GitHub Repository to get the pre-trained CenterPoint-Voxel model.

3 Planned Follow-up

I plan to firstly implement the View-Aware Transformer and get the LLM to make the structure of LiDAR-LLM complete.

After that i want to choose one experiment in the article to conducte, like the 3D Captioning/3D Grounding. But considerd that the dataset may have to be prodced by myself and the trainning of the model need A100 GPUs, it may be hard to explore further.