포인트클라우드와대규모언어모델을결합한 3차원다중모달이해접근방식분류및분석

손석빈, 박수현*, 김중헌

고려대학교, 숙명여자대학교*

lydiasb@korea.ac.kr, soohyun.park@sookmyung.ac.kr, joongheon@korea.ac.kr

A Taxonomy and Analysis of 3D Multimodal Understanding Approaches Integrating Point Clouds and Large Language Models

Seok Bin Son, Soohyun Park*, Joongheon Kim

Korea Univ., Sookmyung Women's Univ.*

요 약

최근 3차원 장면 이해를 고도화하기 위해 포인트 클라우드와 대규모 언어 모델(LLM)을 결합하려는 연구가 활발히 진행되고 있다. 본 논문은 기존 연구들을 처리 전략과 정보 흐름의 관점에서 네 가지 대표적인 패러다임으로 분류하고, 각 접근 방식의 구조적 특징과 설계상의 trade-off를 분석한다. 이를 통해 현재 접근 방식의 한계를 정리하고, 향후 효율적이고 의미적으로 풍부한 3차원 다중모달 추론을 위한 연구 방향을 제시한다.

Ⅰ. 서 론

자율주행, 로보틱스, 혼합현실과 같은 응용 분야의 발전과 함께 3차원 환경을 정밀하게 이해하고 추론하는 기술의 중요성이 지속적으로 증가하고 있다 [1]. 이러한 3차원 환경 이해를 위해 다양한 데이터 표현 방식이 활용되어 왔으며, 그중 포인트 클라우드는 LiDAR 및 RGB-D 센서를 통해 직접 획득 가능하고 공간 기하 정보를 정밀하게 보존할 수 있다는 점에서 대표적인 3차원 표현 방식으로 자리 잡고 있다. 그러나 포인트 클라우드는 비정형적이며 순서가 없는 구조를 가지므로, 기존의 격자 기반 신경망이나 표준 Transformer 구조를 그대로 적용하는 데 근본적인 제약이 존재한다. 한편, 대규모 언어 모델(Large Language Models, LLMs)은 언어 이해와 추론 능력에서 비약적인 성능 향상을 달성하며 다양한 영역으로 확장되고 있다. 특히 최근에는 이미지와 텍스트를 결합한 다중모달 LLM이 2차원 비전 - 언어 태스크에서 높은 성과를 보이며, 이러한 언어 중심 추론 능력을 3차원 공간 이해로 확장하려는 연구가 활발히 진행되고 있다. 그러나 2차원 이미지와 달리 포인트 클라우드는 규칙적인 구조를 가지지 않으며, 3차원 장면 이해를 위해서는 기하 정보와 의미 정보의 결합이 필수적이라는 점에서 새로운 설계상의 어려움을 내포한다. 이러한 배경 하에서 최근 연구들은 포인트 클라우드와 LLM을 효과적으로 연결하기 위한 다양한 구조적 접근 방식을 제안하고 있다. 그러나 기존 연구들은 서로 다른 설계 목표와 가정을 기반으로 전개되어 전체적인 연구 흐름을 체계적으로 파악하는 데 한계가 존재한다. 이에 본 논문은 그림 1과 같이 구성된 포인트 클라우드 기반 다중모달 LLM에 대한 연구들을 공통된 기준에 따라 분류하고, 각 접근 방식의 특징과 한계를 분석함으로써 향후 3차원 다중모달 지능 연구의 방향성을 정리하고자 한다.

Ⅱ. 제안 방법 및 학습 결과 분석

본 논문은 기존 연구들을 처리 전략과 정보 흐름의 관점에서 네 가지 주요 패러다임으로 분류한다. 각 패러다임은 포인트 클라우드와 언어 모델을 연결하는 방식에서 서로 다른 설계 철학을 반영하며, 기하 정확도, 의미적 풍부함, 계산 효율성 사이의 상이한 균형점을 가진다. 다음 절에서는 이러한 네 가지 패러다임을 각각 상세히 분석하고, 각 접근 방식의 구조적 특징과 한계를 구체적으로 논의한다.

그림 1 3차원 멀티모달 LLM 모델

2.1 직접 포인트 클라우드 인코딩 기반 접근

직접 포인트 클라우드 인코딩 기반 접근은 원시 포인트 클라우드를 별도의 중간 표현 없이 3차원 인코더에 직접 입력하여 처리하는 방식을 따른다 [2]. 이 접근에서는 XYZ 좌표와 RGB 색상 정보가 결합된 포인트 집합을 입력으로 사용하며, 기하 구조를 최대한 보존하는 것을 주요 설계 목표로 한다. 이를 위해 PointNet++ 또는 sparse convolution 기반 네트워크와 같은 3차원 전용 인코더가 활용된다. 이러한 방식에서는 포인트 수가 많아질수록 계산 복잡도와 메모리 사용량이 급격히 증가하므로, 전처리 단계에서 좌표 정규화, farthest point sampling, 복셀화와 같은 기법이 적용된다. 이후 다중 스케일 특징 추출을 통해 지역적·전역적 기하 정보를 동시에 인코딩하며, 최종적으로 LLM이 처리 가능한 토큰 시퀀스로 변환한다. 이 과정은 포인트 클라우드의 공간 구조를 언어 모델의 추론 공간으

로 직접 연결한다는 점에서 높은 기하 충실도를 제공한다. 그러나 이러한 접근은 대규모 장면이나 고밀도 포인트 클라우드에 적용할 경우 확장성 문제가 발생하며, LLM의 입력 길이 제한에 의해 토큰 수를 제한해야 한다는 제약을 가진다. 따라서 직접 인코딩 방식은 정밀한 공간 추론이 요구되는 태스크에 적합하나, 계산 효율성 측면에서는 한계를 가진다.

2.2 포인트 클라우드 기반 다중모달 정렬 접근

포인트 클라우드 기반 다중모달 정렬 접근은 3차원 데이터와 텍스트, 이미지, 오디오, 비디오 등 다양한 모달리티를 공통 잠재 공간으로 매핑하는 것을 목표로 한다 [3]. 이 접근에서는 포인트 클라우드를 고차원 의미 임베딩으로 변환한 후, 다른 모달리티의 임베딩과 대조 학습을 통해 의미적으로 정렬한다. 이러한 구조는 특정 모달리티에 국한되지 않는 유연한 추론을 가능하게 하며, 텍스트 질의에 기반한 3차원 객체 검색이나 크로스모달 매칭과 같은 태스크에 효과적으로 활용된다. 포인트 클라우드는 이 과정에서 주로 Transformer 기반 인코더를 통해 처리되며, 기하 구조보다는 의미적 표현에 초점을 둔다. 그러나 다중모달 정렬 과정에서는 포인트 클라우드의 세밀한 기하 정보가 고수준 의미 표현으로 압축되므로, 정확한 공간 관계 추론이나 미세한 구조 분석에는 한계가 존재한다. 따라서 이 접근은 의미 중심의 추론과 상호작용에는 적합하나, 기하 정확도가 중요한 응용에는 제한적이다.

2.3 의미 정보 결합 기반 업샘플링 접근

의미 정보 결합 기반 업샘플링 접근은 희소한 포인트 클라우드를 고해상도로 복원하는 문제를 대상으로 한다 [4]. 이 접근에서는 포인트 클라우드에서 추출한 기하적 특징과 LLM이 생성한 언어 기반 의미 정보를 결합하여, 구조적 완성도와 의미적 일관성을 동시에 향상시키는 것을 목표로 한다. 구체적으로, 포인트 클라우드로부터 다중 스케일 기하 특징을 추출한 후, LLM을 활용하여 장면에 대한 텍스트 설명을 생성한다. 이후 기하 특징과 의미 임베딩을 공통 표현 공간에서 정렬하고 결합하여 업샘플링 네트워크에 입력한다. 이를 통해 단순한 보간 방식이 아닌, 의미적으로 타당한 고밀도 포인트 클라우드 생성을 가능하게 한다. 다만 이 접근은 업샘플링 품질이 LLM이 생성하는 의미 정보의 정확성에 크게 의존하며, 의미 생성 오류가 기하 복원 오류로 이어질 수 있다는 한계를 가진다. 따라서 의미 정보의 신뢰성과 정합성이 핵심적인 요소로 작용한다.

2.4 포인트 클라우드 기반 다중모달 정렬 접근

다중 시점 이미지 기반 접근은 포인트 클라우드를 직접 처리하지 않고, 다중 시점 RGB 이미지와 깊이 정보를 활용하여 3차원 인식을 수행한다 [5]. 이 방식은 기존의 대규모 2차원 비전 - 언어 모델을 재활용할 수 있다는 점에서 계산 효율성과 구현 용이성을 가진다. 이 접근에서는 각 이미지 패치에 대응되는 3차원 좌표를 계산하여 3차원 위치 정보를 시각 토큰에 부여한다. 이후 이러한 3차원 인지 토큰을 LLM 기반 다중모달 모델에 입력하여 공간 추론을 수행한다. 이를 통해 2차원 기반 모델을 확장하여 3차원 객체 위치 추정이나 공간 관계 이해를 가능하게 한다. 그러나 다중 시점 이미지 기반 접근은 정확한 카메라 파라미터와 충분한 시점 다양성이 요구되며, 입력 품질에 따라 성능이 크게 좌우된다. 또한 포인트 클라우드가 제공하는 직접적인 기하 정보에 비해 표현의 정밀도가 제한될 수 있다.

Ⅲ. 결론

본 논문은 포인트 클라우드와 대규모 언어 모델을 결합한 최근 연구들을 네 가지 주요 패러다임으로 분류하고, 각 접근 방식의 구조적 특징과 설계상의 trade-off를 분석하였다. 직접 포인트 클라우드 인코딩 기반 접근은 높은 기하 충실도를 제공하나 확장성에 제약이 있음을 확인하였다. 포인트 클라우드 기반 다중모달 정렬 접근은 유연한 크로스모달 추론을 가능하게 하나 세밀한 공간 정보가 희석됨을 확인하였다. 의미 정보 결합 기반 업샘플링 접근은 재구성 품질을 향상시키지만 의미 생성의 정확성에 의존함을 확인하였다. 다중 시점 이미지 기반 접근은 기존 2차원 모델을 효과적으로 활용할 수 있으나 입력 조건에 민감함을 확인하였다. 이러한 분석은 포인트 클라우드 - LLM 통합 연구에서 단일 접근법만으로는 기하 정확성, 의미적 풍부함, 계산 효율성을 동시에 만족시키기 어렵다는 점을 시사한다. 향후 연구에서는 이러한 패러다임을 융합한 하이브리드 구조 설계와 함께, 공정한 비교를 위한 통합 벤치마크 및 대규모 다중모달 3차원 데이터셋 구축이 중요해질 것으로 판단한다.

ACKNOWLEDGMENT

이 성과는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(RS-2025-00561377). 본 논문의 교신저자는 김중헌임.

참 고 문 헌

[1] A. Memon et al., “Prior-free 3d human pose estimation in a video using limb-vectors,” *ICT Express*, vol. 10, no. 6, pp. 1266 - 1272, December 2024.

[2] S. Chen et al., “LL3DA: visual interactive instruction tuning for omni-3d understanding, reasoning, and planning,” in P*roc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, (CVPR)*, Seattle, WA, USA, June 2024, pp. 26418 - 26428.

[3] H. Li et al., “Point-bind & Point-LLM: Aligning point cloud with multi- modality for 3d understanding, generation, and instruction following,” *arXiv preprint arXiv:2309.00615*, 2023.

[4] Z. Zhang et al., “PULLM: A multimodal framework for enhanced 3d point cloud upsampling using large language models,” in *Proc. of the ACM/SIGAPP Symposium on Applied Computing, (SAC)*, Catania International Airport, Catania, Italy, April 2025, pp. 1223 - 1230.

[5] C. Zhu et al., “Llava-3d: A simple yet effective pathway to empowering LMMs with 3d-awareness,” *arXiv preprint arXiv:2409.18125*, 2024.