LLM의 공간 정보 이해를 위한 텍스트 기반 벡터 표현 압축에 관한 연구

박종현, 김예찬, 김성헌, 주종민, 이민기, 전문구
광주과학기술원

{citizen135, yechankim, ha307702, luck2u99, leemingi}@gm.gist.ac.kr, mgjeon@gist.ac.kr

A Study on Text-Based Vector Representation Compression for Spatial Information Understanding in Large Language Models

Jong Hyun Park, Sungheon Kim, Jongmin Joo, Mingi Lee, Moongu jeon
Gwangju Institute of Science and Technology.

요 약

지도나 범례가 포함된 공간 이미지에 대한 비전언어모델의 공간 이해에는 여전히 한계가 있어, 지형 이해가 필요한 AI Agent 환경에서는 텍스트 기반 공간 표현이 활용되고 있다. 그러나 SVG와 같은 텍스트 기반 벡터 지형 데이터는 객체 수 증가에 따라 LLM의 컨텍스트 길이 부담이 커지는 문제가 있다. 본 논문에서는 인접 폴리곤을 통합하고 Ramer-Douglas- Peucker(RDP) 알고리즘을 적용하여 공간적 의미를 유지하면서 텍스트 기반 지형 데이터를 압축하는 방법을 제안한다. 이를 통해 고수준 공간적 특성을 보존하면서도 LLM 입력 길이를 효과적으로 감소시키는 방법은 제안한다.

Ⅰ. 서 론

최근 대규모언어모델(Large Language Model, LLM)의 성능이 빠르게 발전하고 있다. ChatGPT 의 첫 등장 당시와 비교하면 환각과 같은 문제가 일부 개선되고 있으며 음성, 이미지, 영상 등을 입력으로 받는 멀티모달 언어 모델들도 등장하고 있다.

이미지를 입력으로 받는 비전언어모델(Vision-Language Model, VLM)은 등장 초반에는 이미지와 함께 주어지는 텍스트 프롬프트에 크게 의존하는 문제가 있어, 일부 퍼즐과 같은 문제에서 이미지를 아예 넣지 않거나 노이즈 이미지를 넣는 것이 좋을 만큼 낮은 영상 이해 능력을 보였다[1]. 이후 VLM 은 빠른 속도로 발전하여 공간 이해 능력은 GPT-5 에서 큰 향상을 보였지만 여전히 인간 평가자와 비교하면 낮은 성능을 보인다[2]. 특히 자연영상보다 범례가 있는 지도와 같은 이미지에 대한 공간 이해 능력은 더욱 낮은 성능을 보이며 이는 상업용 모델과 오픈 소스 모델 비교에서도 큰 격차를 보인다[3][4].

LLM 의 추론 능력을 이용한 AI Agent 는 다양한 분야에서 활용되고 있다. 도시계획, 국방, 감시나 환경과 상호작용이 필요한 경우 LLM 의 공간 이해 능력이 필수적이다. 이러한 환경에서의 VLM을 사용한 AI Agent 연구는 앞서 설명한 이유로 낮은 성능을 보여 지형을 픽셀 단위로 텍스트로 표현하여 LLM 을 이용해서 실험을 진행한다[6][7]. 간단한 지형과 공간이해가 필요한 AI Agent 연구에서는 픽셀 단위의 지형표현으로 연구가 어느정도 가능하지만 대규모 지형 이해 및 입력을 필요로 하는 연구에는 한계가 있다.

기존에는 이러한 지형 데이터를 텍스트로 표현할 때, 공간 참조 시스템(Spatial Reference System)에서 WKT(Well-Known Text)를 사용한다. WKT 는 지도 상의 형상(포인트, 선, 폴리곤, 다각형 등)을 구성하는 좌표들을 텍스트로 표현한다. 이러한 표현 형태에 대해 LLM 의 이해 능력의 관한 연구는 존재한다[8]. 하지만 VLM 의 이미지에 대한 공간 이해와 WKT 직접적인 비교는 힘들다. 텍스트로 표현된 객체들 간의 거리와 위치 관계는 텍스트 상의 위치와 연관이 없으며, 관계(겹침, 닿음, 떨어짐 등)과 같은 공간관계(Spatial relation)를 별도의 Geospatial topology 로 표현하여 관계를 알아 낼 수 있다. 이미지가 표현하는 범위가 크지만 관심 범위가 정해져 있는 경우 해당 범위를 크롭하면 되지만, 텍스트 표현은 객체가 많을수록 검색해야 하는 범위가 커진다. 따라서 LLM 의 부담을 줄이고 제한된 컨텍스트 크기를 가진 LLM 을 위해서 WKT, SVG(Scalable Vector Graphics)와 같은 텍스트 기반 벡터 표현 데이터를 압축하여 사용하는 변환 방법에 대해 소개한다.

본 논문에서는 지형 이해가 필요한 AI Agent 에 활용할 수 있도록 LLM 컨텍스트 크기에 맞게 지형 데이터를 압축하는 것을 다룬다. 인접한 폴리곤을 통합하고 폴리곤 표현을 단순화하고 관심 영역을 찾은 경우, 해당 지역에 대해 크롭하는 것을 다룬다.

Ⅱ. 본론

본논문에서 예시로 사용하는 데이터는 Arma 3 게임의 맵 중 Altis 의 SVG 데이터를 LLM 이 공간 정보를 이해할 수 있도록 압축하는 과정에 대해서 다룬다. 해당 데이터에는 총 10 개의 레이어(지형, 숲, 암석, 등고선, 도로, 객체 등)가 있으며, 그 중 1 개 레이어(구조물)만을 압축하여 시각화 하였다[그림 1].

그림1. 압축 순서에 따른 지형 데이터의 변화. (좌측) 10개의 레이어가 포함된 공간 데이터. (중간) 10개의 레이어 중 1개 레이어(구조물)만을 추출하고 인접한 레이어 통합. (우측) 다각형 근사 알고리즘을 통해 각 폴리곤 표현에 필요한 포인트 수를 줄임.

압축은 두 과정으로 나누어진다. 먼저 Step 1 에서는 전체 레이어에서 관심 레이어를 추출하고 인접 폴리곤을 통합함으로서 관련 없는 레이어끼리 통합되는 것을 막고 LLM 컨텍스트 크기에 대한 부담을 줄인다. 인접 폴리곤은 크기 30 의 패딩을 두어 겹치는 폴리곤을 통합하여, 구조물, 건물이 많은 도심지의 경우에는 하나의 통합된 폴리곤으로 LLM 이 해당 구역이 건물이 많은 도심지로 이해할 수 있도록 했다.

Step 2 에서는 다각형 근사 알고리즘 중 Ramer-Douglas-Perucker(RDP)를 적용하여 통합된 폴리곤의 형상을 유지하면서 불필요한 정점을 제거하는 방식으로 추가적인 압축을 수행한다. RDP 알고리즘의 핵심 파라미터는 ε (epsilon)으로, 이는 원본 다각형과 근사된 다각형 사이에서 허용되는 최대 거리 오차를 의미한다. [그림 1]에서는 ε 값을 10 으로 설정하였으며, 이는 SVG 좌표계 상에서 개별 구조물의 세부적인 외곽선 변형은 허용하되, 건물 군집의 전반적인 공간적 배치와 형태는 유지하도록 하기 위함이다. 해당 설정을 통해 LLM 입력 관점에서는 세밀한 기하 정보로 인한 컨텍스트 길이 증가를 억제하면서도, 특정 영역이 구조물이 밀집된 지역인지 여부와 같은 공간적 의미는 손실되지 않도록 하였다. 결과적으로 Step 1 에서 인접한 폴리곤을 통합함으로써 복잡해진 폴리곤 형상을 단순화하고 컨텍스트 길이는 줄이면서 LLM 이 고수준의 공간적 특성을 이해하는 데 적합한 표현을 제공한다.

컨텍스트 크기나 관심 영역이 정해져 있는 경우에는 이러한 압축 방법을 생략하거나 파라미터를 조정해 사용할 수 있다.

Ⅲ. 결론

서론에서는 LLM 의 공간 이해에 대해 지도 이미지와 공간의 텍스트 표현 형태를 비교하며 아직까지는 텍스트 기반 벡터 표현 데이터의 공간 정보 이해가 LLM 에 더 적합하며, 본론에서는 텍스트 기반 벡터 표현 데이터 중 하나인 SVG 파일을 LLM 의 이해를 돕기 위해 압축하는 방법에 대해 설명했다.

지도나 범례가 있는 공간 정보 이미지에 대해 VLM 이 기존의 공간 정보 이해의 한계를 극복한 모델이 등장하기까지 시간이 오래 걸릴 것 같지는 않다. 그럼에도 이미지를 입력으로 받는 VLM 까지 사용할 필요가 없거나 정해진 형태의 지형 데이터를 이용하는 환경에서는 무겁거나 비싼 모델을 사용할 필요 없이 텍스트 기반 벡터 표현 데이터를 압축하여 오픈소스 소규모언어모델(sLM)을 그대로 사용하거나 미세조정하여 모델이 공간에 대한 이해를 할 수 있도록 할 수 있다.

추후 연구에서는 컨텍스트가 얼마나 압축되는지, 그에 따라 LLM 의 이해도가 얼마나 더 커지는지 비교해볼 필요가 있다.

ACKNOWLEDGMENT

이 성과는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (No. 2023R1A2C2006264).

참 고 문 헌

[1] Wang, Jiayu, et al. "Is a picture worth a thousand words? delving into spatial reasoning for vision language models." *Advances in Neural Information Processing Systems* 37 (2024): 75392-75421.

[2] Cai, Zhongang, et al. "Has gpt-5 achieved spatial intelligence? an empirical study." *arXiv preprint arXiv:2508.13142* 3 (2025).

[3] Pyo, Jiyoon, et al. "FRIEDA: Benchmarking Multi-Step Cartographic Reasoning in Vision-Language Models." *arXiv preprint arXiv:2512.08016* (2025).

[4] Srivastava, Varun, et al. "MapIQ: Evaluating multimodal large language models for map question answering." *arXiv preprint arXiv:2507.11625* (2025).

[6] Anne, Timothée, et al. "Harnessing language for coordination: A framework and benchmark for llm-driven multi-agent control." *IEEE Transactions on Games* (2025).

[7] Nasir, Muhammad Umair, Steven James, and Julian Togelius. "Gametraversalbenchmark: Evaluating planning abilities of large language models through traversing 2d game maps. "*Advances in Neural Information Processing Systems* 37 (2024): 31813-31827.

[8] Ji, Yuhan, et al. "Foundation models for geospatial reasoning: assessing the capabilities of large language models in understanding geometries and topological spatial relations." *International Journal of Geographical Information Science* (2025): 1-38.