Reasoning Segmentation 의 실패 유형 분석

윤경윤 [1], 조영준 [2]
[1] 전남대학교 인공지능융합학과 석사과정
[2] 전남대학교 인공지능융합학과 교수

kyungyoon201@gmail.com, yj.cho@jnu.ac.kr

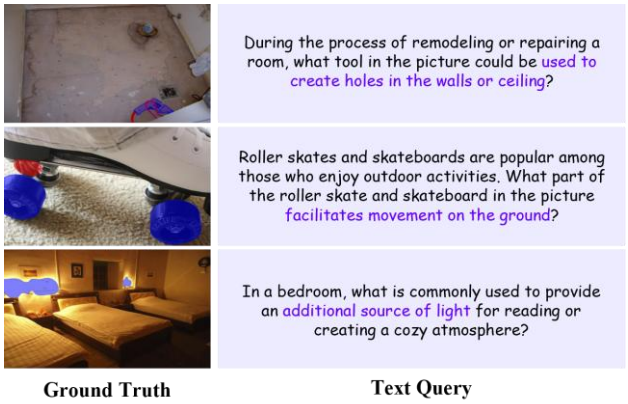
Analysis of Failure Types in Reasoning Segmentation

Yoon Kyung-yoon[1], Cho Yeong-jun [1]
[1]Dept. of Artificial Intelligence Convergence, Chonnam National University

요 약

Reasoning Segmentation 은 비전-언어 모델(VLM)과 세분화 모델을 결합하여 간접적인 질의에 대응하는 영역을 마스크로 예측하는 새로운 세분화 과제이다. 최근 다양한 접근법이 제안되었으나, 실제 성능은 여전히 불안정하며 특정 상황에서 성능 저하가 빈번히 발생한다. 본 연구에서는 주요 Reasoning Segmentation 모델의 추론 및 분할 결과를 체계적으로 분석하여 성능 저하를 일으키는 실패 양상을 분류하였다. 이를 통해 향후 모델 개선 방향성을 제시하고자 한다.

1. 서론



(그림 1) Reasoning Segmentation 의 예시.

기존의 세분화 연구는 명확하게 정의된 객체를 탐지하는 데 집중해왔다. 그러나 실제 응용 환경에서는 단순히 객체를 식별하는 것만으로는 충분하지 않은 경우가 많다. 예를 들어, “책상 위에서 빛을 반사하는 부분을 찾아라” 와 같이 간접적이거나 맥락에 의존하는 질의에 대응하기 위해서는 단순한 물체 검출 능력을 넘어선 추론 능력이 필요하다. 이러한 요구에서 출발한 것이 바로 Reasoning Segmentation 이다.

Reasoning Segmentation 은 VLM 의 추론 능력을 활용하여 텍스트 질의가 가리키는 대상의 의미적 위치를 먼저 추정한 후, 세분화 모델을 통해 해당 위치를 마스크로 변환하는 과정을 거친다. 이 과제는 Referring Segmentation 과 달리 질의가 직접적으로 객체의 이름을 언급하지 않고, 맥락적 · 관계적 정보를 통해 대상을 지칭하는 경우가 많기 때문에 더 높은 수준의 언어 이해와 시각적 추론을 요구한다.

최근 몇 년간 LISA[1], READ[2], SESAME[3] 등 다양한 연구들이 Reasoning Segmentation 을 제안하고 성능을 개선하고자 했지만, 여전히 실제 응용에서는 여러 한계가 존재한다. 특히 VLM 의 추론 결과와 세분화 모델의 출력 간의 불일치, 세분화 경계 품질 저하, 질의 자체의 복잡성에 기인한 오해석 등이 주요 문제로 보고되고 있다. 따라서 단순히 모델의 정량적 성능을 높이는 것에서 나아가, 실제로 어떠한 실패 양상이 반복적으로 나타나는지를 분석하고 정리하는 작업은 학문적으로나 실용적으로 중요한 의미를 가진다.

본 연구는 ReasonSeg 벤치마크를 활용하여 주요 Reasoning Segmentation 모델(LISA, READ)의 결과를 면밀히 분석하였다. 그리고 성능 저하를 유발하는 실패 사례를 유형별로 분류함으로써, 향후 연구가 집중해야 할 개선 지점을 제시하고자 한다.

2. 관련 연구

Reasoning Segmentation 은 LISA[1]에서 처음 체계적으로 제안되었으며, 이후 여러 후속 연구가 등장하였다. LISA 는 LLaVA[4]와 같은 VLM 에서 추출한 <SEG> 토큰을 SAM(Segment Anything Model)[5]에 직

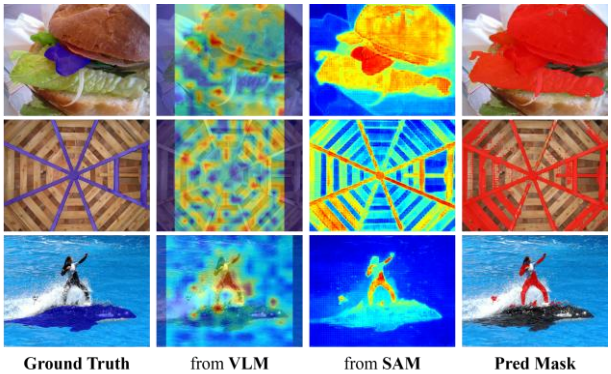
접 전달하는 방식을 사용하여 초기 성능을 입증하였다. READ 는 VLM 이 생성한 유사도 맵으로부터 포인트 프롬프트를 추출하여 SAM 에 제공함으로써 더 정교한 위치 정보를 전달하고자 했다. 또한 SESAME 은 질의가 허위 전제를 포함할 수 있는 상황까지 고려하여 모델의 추론적 견고성을 강화하는 방향으로 확장되었다.

이와 같은 Reasoning Segmentation 연구는 비전-언어 모델의 성능에 크게 의존한다. 최근 VLM, 예를 들어 LLaVA, BLIP-2[6], Flamingo[7] 등은 이미지와 텍스트 간의 복합적 연관성을 학습함으로써 복잡한 질의에 대한 추론 능력을 확보하고 있다. 그러나 이러한 모델들은 일반적인 시각 언어 이해에는 강점을 보이지만, 공간적 정밀도나 마스크 수준의 세분화 품질을 보장하지는 못한다. 반대로 SAM 과 같은 세분화 모델은 다양한 시각적 프롬프트를 기반으로 정밀한 마스크를 생성할 수 있으나, 언어적 맥락을 직접 해석하는 데에는 취약하다.

즉, Reasoning Segmentation 은 언어와 시각이라는 두 모듈의 강점을 결합하는 과정에서 발생하는 불일치를 어떻게 해소할 것인가라는 근본적인 문제에 직면해 있다. 따라서 기존 연구의 성과와 한계를 면밀히 검토함과 동시에, 실패 사례를 체계적으로 분류하고 원인을 규명하는 연구가 필요하다.

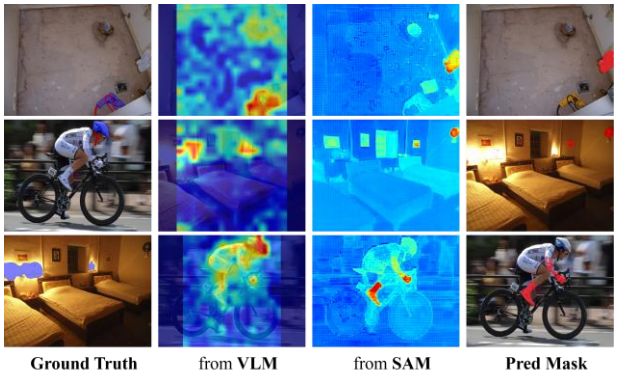
3. Reasoning Segmentation 실패 유형

본 연구에서는 ReasonSeg 벤치마크의 검증(val) 200 장과 시험(test) 779 장을 대상으로 LISA 와 READ 모델의 출력을 분석하였다. 분석 절차는 먼저 각 질의에 대해 VLM 이 생성한 활성 맵을 수집하고, 이어서 세분화 모델이 생성한 마스크 출력을 비교하는 방식으로 이루어졌다. 이 과정을 통해 모델이 질의 해석 단계에서부터 최종 마스크 생성 단계에 이르기까지 어떠한 오류가 발생하는지를 정성적으로 파악할 수 있었다.



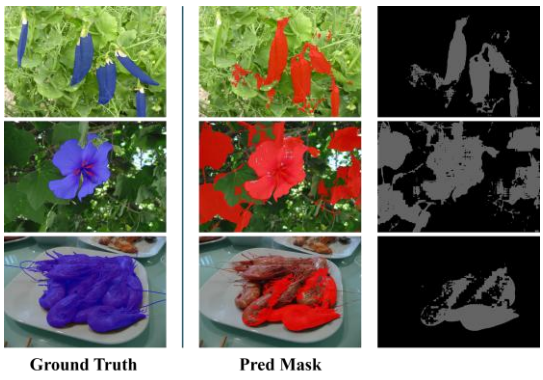
(그림 2) VLM 이 질의의 의미를 잘못 해석한 사례.

분석 결과 실패 사례는 크게 세 가지 유형으로 분류되었다. 첫 번째 유형은 VLM 이 질의의 의미를 올바르게 해석하지 못하는 경우이다. 예컨대 “음식에서 단백질이 많이 포함된 부분”이라는 질의에 대해 특정 부위를 정확히 찾아야 하지만, 모델이 음식 전체를 활성화하거나 전혀 다른 영역을 활성화하는 사례가 관찰되었다. 이는 언어적 맥락 이해 부족, 또는 복합적 속성을 가진 질의에 대한 추론 실패로 볼 수 있다.



(그림 3) 세분화 모델에서 잘못된 위치를 활성화하는 사례.

두 번째 유형은 VLM 은 비교적 올바른 위치를 지정했으나, 세분화 모델이 이를 잘못된 마스크로 변환하는 경우이다. 예를 들어 “벽에 못을 박을 때 사용하는 것”이라는 질의에서 VLM 은 드릴 부위에 집중했지만, 세분화 모델은 드릴과 무관한 주변 객체를 마스크로 출력하는 사례가 이에 해당한다. 이러한 경우는 VLM 과 세분화 모델 간 구조적 불일치가 원인으로 작용한다.



(그림 4) 마스크의 품질이 좋지 않은 사례.

세 번째 유형은 위치 자체는 대체로 맞지만, 생성된 마스크의 품질이 낮은 경우이다. 구체적으로는 객체 경계가 불명확하거나 구멍(hole)이 많이 발생하거나, 불필요한 배경이 포함된 경우이다. 예컨대 “식물에서 씨를 포함하는 부분”을 찾는 질의에서 열매뿐만 아니라 줄기와 잎까지 함께 포함된 마스크가 출력되는 사례가 대표적이다. 이는 세분화 모델의 정밀한

경계 예측 한계와 정제 과정의 부재가 주요 원인으로 지적된다.

4. 결과 및 논의

분석 결과, 실패 사례 가운데 가장 두드러진 유형은 마스크 품질 저하와 관련된 것이었다. 이어서 VLM 단계에서 질의를 잘못 해석하거나 부적절한 객체에 집중하는 사례가 두 번째로 많이 관찰되었으며, 세분화 모델이 잘못된 위치를 활성화하는 경우도 적지 않게 나타났다. 이러한 결과는 Reasoning Segmentation 의 한계가 단일 모듈의 문제라기보다, 추론과 세분화 양쪽에서 모두 발생할 수 있음을 보여준다. 이를 통해 몇 가지 시사점을 얻을 수 있다. 첫째, VLM 이 질의를 보다 정확하게 이해하고 이미지 내 적절한 대상을 식별하도록 추론 능력을 강화할 필요가 있다. 이를 위해 다양한 상황을 반영하는 데이터셋 구축이나 학습 전략이 요구된다. 둘째, VLM 의 출력과 세분화 모델을 연결하는 과정에서 발생하는 구조적 불일치를 완화할 수 있는 개선이 필요하다. 예컨대, VLM 의 텍스트 임베딩을 효과적으로 이해하고 처리할 수 있는 세분화 모델의 설계가 가능하다. 셋째, 마스크 품질 향상을 위해 노이즈 제거와 경계 정제 기법을 도입함으로써 실제 응용 환경에서 신뢰할 수 있는 결과를 제공해야 한다.

5. 결론

본 연구에서는 Reasoning Segmentation 모델(LISA, READ)을 대상으로 성능 저하를 유발하는 실패 유형을 분류 · 분석하였다. 분석 결과는 세 가지로 요약된다. 첫째, VLM 이 질의의 의미를 잘못 해석하여 전혀 다른 영역을 지정하는 경우가 존재한다. 둘째, VLM 은 대체로 올바른 위치를 활성화했으나 세분화 모델이 전혀 다른 마스크를 생성하는 경우가 확인되었다. 셋째, 위치는 알맞지만 경계가 불명확하거나 불필요한 부분을 포함하는 마스크 품질 저하 문제가 빈번히 발생했다.

이러한 결과를 종합하면, Reasoning Segmentation 의 성능 향상을 위해서는 세 가지 개선 방향이 필요하다. 첫째, 언어적 추론 능력을 강화하여 질의 해석 정확도를 높여야 한다. 둘째, VLM 과 세분화 모델 간의 구조을 개선하여 추론 결과가 세분화 단계로 자연스럽게 이어지도록 해야 한다. 셋째, 마스크 품질 향상을 위한 정제 기법을 도입하여 실제 응용에서 신뢰할 수 있는 결과를 제공해야 한다.

본 연구는 단순히 성능 지표를 제시하는 것에 그치지 않고, 실패 사례를 유형별로 분류하여 구체적으로 제시함으로써 Reasoning Segmentation 연구가 나아가야 할 방향성을 확인했다는 데 의의가 있다. 향후 연구는 이러한 실패 모드를 극복하는 방법론을 중심으로 발전할 것이며, 이는 Reasoning Segmentation 의 실질적 응용 가능성을 크게 확장하는 데 기여할 것이다.

6. 한계점

본 연구는 ReasonSeg 벤치마크를 기반으로 대표적인 Reasoning Segmentation 모델인 LISA 와 READ 를 대상으로 사례 분석을 수행하였다. 그러나 분석의 범위가 일부 모델에 국한되어 있어, 다른 최신 모델이나 변형 아키텍처에 대해서는 동일한 결론을 일반화하기 어렵다는 한계가 있다. 또한 본 연구의 분석은 주로 정성적 비교에 초점을 맞추었기 때문에, 보다 엄밀한 정량적 통계와 지표를 활용한 체계적 검증이 부족하다. 이러한 제약에도 불구하고, 본 연구는 실패 양상을 분류하고 원인을 규명하는 출발점을 제시했다는 점에서 의미가 있다.

7. 향후 연구

향후 연구에서는 ReasonSeg 외의 다양한 데이터셋을 포함하여 모델의 일반화 가능성을 평가할 필요가 있다. 특히 복잡한 텍스트 질의나 멀티 모달 맥락을 반영하는 새로운 벤치마크를 구축하는 것이 유용할 것이다. 또한 VLM 과 세분화 모델을 단순히 연결하는 수준을 넘어, 두 모듈을 통합적으로 학습할 수 있는 end-to-end 구조를 설계하는 방향이 요구된다. 아울러 마스크 품질 문제를 해결하기 위해 경계 정제(refinement) 기법이나 후처리 모듈을 접목하는 방법 역시 중요한 연구 과제가 될 수 있다. 이러한 시도들은 Reasoning Segmentation 의 신뢰도를 높이고 실제 응용 환경에서 활용성을 확장하는 데 기여할 것으로 기대된다.

8. 감사의 글

이 논문은 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원-대학 ICT 연구센터(ITRC)의 지원을 받아 수행된 연구임(IITP-2025-RS-2024-00437718). 또한 본 결과물은 농림축산식품부의 재원으로 농림식품기술기획평가원의 농식품과학기술융합형연구인력양성사업의 지원을 받아 연구되었음(RS-2024-00397026). 아울러 본 연구는 과학기술정보통신부 및 정보통신기획평가원의 인공지능융합혁신인재양성사업 연구 결과로 수행되었음(IITP-2023-RS-2023-00256629).

참고문헌

[1] Lai, X. et al. “LISA: Reasoning Segmentation via Large Language Model,” CVPR, 2024.

[2] Qian, R. et al. “READ: Reasoning to Attend,” CVPR, 2025.

[3] Wu, T.-H. et al. “SESAME: Teaching LMMs to Overcome False Premises,” CVPR, 2024.

[4] Liu, H. et al. “LLaVA: Visual Instruction Tuning,” NeurIPS, 2023.

[5] Kirillov, A. et al. “Segment Anything,” ICCV, 2023.

[6] Li, J., Li, D., Savarese, S., and Hoi, S. “BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models,” ICML, 2023.

[7] Alayrac, J.-B. et al. “Flamingo: a Visual Language Model for Few-Shot Learning,” NeurIPS, 2022.