대규모 비전-언어 모델의 환각 완화를 위한 증거 기반 다중 에이전트 추론 프레임워크

김수현, 심병효*
서울대학교

soohyunkim@islab.snu.ac.kr, *bshim@islab.snu.ac.kr

An Evidence-Based Multi-Agent Reasoning Framework for Mitigating Hallucinations in Large Vision-Language Models

Kim Soo Hyun, Shim Byong Hyo*
Seoul National Univ.

요 약

대규모 비전-언어 모델은 이미지 캡셔닝과 시각적 질의응답 등 다양한 응용 분야에서 활용되고 있으나, 입력 이미지와 일치하지 않는 객체나 관계를 생성하는 환각 문제가 여전히 존재한다. 본 논문에서는 이러한 문제를 완화하기 위해 증거 기반 다중 에이전트 추론 프레임워크를 제안한다. 제안 기법은 두 개의 이질적인 비전-언어 에이전트가 독립적으로 응답을 생성하고, 응답 간 불일치가 발생한 경우 외부 비전 도구를 선택적으로 호출하여 시각적 증거를 수집하고 이를 바탕으로 토론을 수행한다. 실험 결과, 제안 기법은 표준 환각 벤치마크에서 객체 및 관계 환각을 효과적으로 감소시키는 것을 확인하였다.

Ⅰ. 서 론

최근 대규모 비전-언어 모델은 이미지와 텍스트를 공동으로 이해하고 생성하는 능력을 바탕으로 이미지 캡셔닝과 시각적 질의응답 등 다양한 응용 분야에서 활용되고 있다[1]. 그러나 이러한 모델은 종종 입력 이미지와 일치하지 않는 객체를 생성하거나, 객체의 속성 및 공간적 관계를 잘못 서술하는 환각 문제가 여전히 존재한다. 이러한 오류는 모델의 출력이 명확한 시각적 근거에 기반하지 않고 언어적 사전 지식에 과도하게 의존할 때 발생한다[2].

기존 연구들은 미세 조정이나 디코딩 전략을 통해 환각을 완화하고자 하였으나[3,4], 추가 학습 비용이 필요하거나 추론 과정에서 생성 결과의 근거를 명확히 설명하기 어렵다는 한계를 가진다. 본 논문에서는 이러한 한계를 극복하기 위해, 서로 다른 관점을 가진 다중 에이전트가 동일한 입력에 대해 독립적으로 응답을 생성하고, 응답 간 불일치가 발생할 경우 이를 중심으로 상호 검증 과정을 수행하는 증거 기반 다중 에이전트 추론 프레임워크를 제안한다.

Ⅱ. 본론

본 논문에서 제안하는 프레임워크는 동일한 이미지를 입력으로 하는 두 개의 이질적인 비전-언어 에이전트를 활용한다. 첫 번째 에이전트는 이미지로부터 객체와 관계를 포함하는 장면 그래프를 생성하고, 두 번째 에이전트는 동일한 이미지에 대한 간결한 캡션을 생성한다. 생성된 장면 그래프와 캡션은 각 에이전트가 질문에 대한 결정을 내리는 데 활용되는 요약 정보로 사용된다. 각 에이전트는 자신이 생성한 요약 정보를 기반으로 질문에 대한 초기 응답을 독립적으로 생성한다.

두 에이전트의 응답이 일치하는 경우, 추가적인 연산 없이 해당 응답을 최종 출력으로 사용한다. 반면, 두 응답이 불일치할 경우 에이전트 간의 토론 과정이 시작된다. 이 과정에서 에이전트들은 필요에 따라 객체 탐지, 분할, 문자 인식, 깊이 추정과 같은 외부 비전 도구를 선택적으로 호출하여 시각적 증거를 수집할 수 있다. 수집된 증거는 두 에이전트 모두에게 공유되며, 각 에이전트는 이를 근거로 자신의 주장을 수정하거나 상대방의 주장에 대해 반박한다. 이러한 과정을 통해 에이전트들은 점진적으로 합의에 도달하며, 환각 가능성이 높은 경우에만 추가적인 검증을 수행함으로써 불필요한 계산 비용을 효과적으로 줄인다.

제안 방법의 성능은 MME-Hallucination 벤치마크를 사용하여 평가하였다. 비교를 위해 기본 모델의 성능과, 토론 단계 없이 각 에이전트가 생성한 장면 그래프 또는 캡션을 요약 정보로만 활용하여 단일 응답을 생성하는 단일 에이전트 설정을 함께 평가하였다. 표에서 확인할 수 있듯이, 제안하는 프레임워크는 두 백본 모두에서 가장 높은 MME-H 총점을 달성하였다. 특히 LLaVA-NeXT 백본에서는 기본 모델 대비 총 점수가 533.3 에서 641.7 로 크게 향상되었다.

Backbone	*Method*	*MME-H Total*
LLaVA-NeXT	Baseline	533.3
	Single Agent	595.0
InternVL2	Baseline	545.0
	Single Agent	596.7
	Proposed Framework	**641.7**

Ⅲ. 결론

본 논문에서는 대규모 비전-언어 모델에서 발생하는 환각 문제를 완화하기 위해, 증거 기반 다중 에이전트 추론 프레임워크를 제안하였다. 제안 기법은 서로 다른 요약 정보를 활용하는 두 에이전트가 독립적으로 응답을 생성하고, 응답 간 불일치가 발생한 경우에만 선택적으로 시각적 증거를 수집하여 상호 검증을 수행하도록 설계되었다. 실험 결과, 제안 프레임워크는 두 백본 모두에서 기존 설정 대비 가장 높은 성능을 달성하였다. 본 연구는 추가 학습 없이 다양한 비전-언어 모델에 적용 가능하다는 점에서 실용적인 장점을 가지며, 향후 비디오 입력이나 보다 다양한 외부 시각 도구로의 확장이 가능할 것으로 기대된다.

참 고 문 헌

[1] Liu H., Li C., Li Y., Li B., Zhang Y., Shen S., and Lee Y. J., “Llava-next:Improved reasoning, ocr, and world knowledge,” January 2024.

[2] Kim J., Kim H., Kim Y., and Ro Y. M. “Code: Contrasting self-generated description to combat hallucination in large multi-modal models,” in Advances in Neural Information Processing Systems (NeurIPS), 2024.

[3] Lyu X., Chen B., Gao L., and Shen H. T. “Alleviating hallucinations in large vision-language models through hallucination-induced optimization,” in Advances in Neural Information Processing Systems (NeurIPS), 2024.

[4] Leng S., Zhang H., Chen G., Li X., Lu S., Miao C., and Bing L. “Mitigating object hallucinations in large vision-language models through visual contrastive decoding,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 13872- 13882.