

ZCD: LVLM 환각을 완화하는 Zero-Image Contrastive Decoding 기법

하지우, 소진현*
대구경북과학기술원

djiwoo20@dgist.ac.kr, *jinhyun@dgist.ac.kr

ZCD: Mitigating hallucinations in Large Vision-Language Models through Zero-image Contrastive Decoding

Jiwoo Ha, Jinhyun So*
DGIST

요약

거대 시각-언어 모델에서 환각을 완화하는 문제는 모델의 신뢰도를 높이기 위해 반드시 해결되어야 할 과제이다. 본 논문에서는, 모든 픽셀 값이 0으로만 채워진 영 이미지와의 대조적 디코딩을 활용한 환각 완화 기법인 ZCD(Zero-image Contrastive Decoding)를 제안한다. LLaVA-1.5 모델을 기반으로 한 실험 결과, ZCD는 베이스라인 및 기존 완화 기법인 VCD 대비 POPE, MME, AMBER 벤치마크에서 더 높은 성능을 보였다. 이를 통해 ZCD가 학습 없이 LVLM의 객체 환각을 효과적으로 억제하는 디코딩 기법임을 입증하였다.

I. 서론

거대 시각-언어 모델(Large Vision-Language Model, LVLM)은 거대 언어 모델의 추론 능력을 활용하여 다양한 시각 과제를 처리할 수 있지만 여전히 환각(Hallucination) 문제에 직면해 있다. 환각은 주어진 시각 입력에 대해 LVLM이 존재하지 않는 객체나 사실과 맞지 않는 관계·속성을 포함한 답변을 생성하는 현상을 뜻한다[5]. 이러한 현상은 의료 영상 분석, 자율주행 등 실제 응용에서 심각한 사고를 야기할 수 있다. 환각을 완화하기 위해 다양한 방법이 제안되었으며, 그중 VCD(Visual Contrastive Decoding) 기법은 학습 없이 환각을 완화하는 유용한 기법이다. VCD는 원본 이미지 입력에 가우시안 노이즈를 더해 부정 로짓(negative logit), 즉 이미지와 관련 없이 언어 편향(language prior)에만 의존한 출력을 생성한다. 이후 이를 원본 로짓과 대조하여 모델이 언어 편향에만 의존하지 않고 시각적 정보를 더 반영한 출력을 생성할 수 있도록 한다[2]. 그러나 가우시안 노이즈가 추가된 이미지에도 여전히 원본 이미지의 정보의 일부가 남아있기 때문에, 완벽히 언어 편향에만 의존한 부정 로짓을 만들 수 없다[9]. 따라서 본 논문에서는 이러한 한계를 해결하기 위해 가우시안 노이즈를 더하는 대신, 원본 이미지에 대한 정보를 완전히 제거한 0으로만 채워진 영 이미지(Zero-Image)와 대조적 디코딩을 하는 ZCD 기법을 제안한다.

II. 배경 지식

θ 로 모델링된 LVLM은 시각 인코더를 통해 시각 입력 v 를 처리하고, 텍스트 프롬프트 x 와 결합하여 이에 대한 응답 y 를 생성한다. y 는 자기회귀적으로(auto-regressive) 생성되며, 각 시점 t 에서 토큰 y_t 는 다음과 같이 확률적으로 선택된다.

$$y_t \sim p_\theta(y_t | v, x, y_{<t}) \propto \exp \text{logit}_\theta(y_t | v, x, y_{<t}) \quad (1)$$

여기서 $y_{<t}$ 는 $(t-1)$ 시점까지 생성된 토큰을 의미한다. $\text{logit}_\theta(\cdot)$ 은 후보 토큰에 대해 모델이 계산한 점수를 의미하며, 이후 *softmax* 함수를 통해 확률 분포로 변환된다. 기존 연구에 따르면, 이러한 디코딩 과정에서 LVLM은 시각 입력 v 보다 거대 언어 모델이 학습한 언어 편향에 의존하는 경향이 있다. 이러한 현상은 LVLM의 객체 환각(object hallucination) 원인 중 하나로 지적되고 있다[2,8].

III. 제안 기법

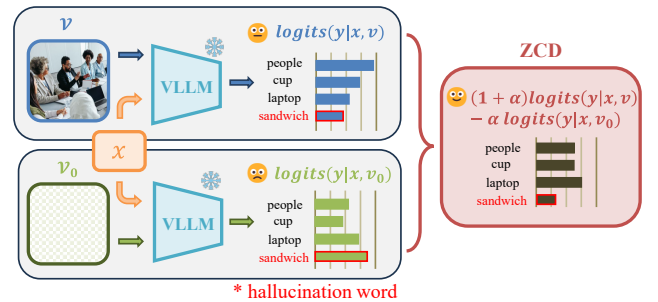


그림 1 ZCD의 전체적인 진행 과정

본 논문에서 제안하는 ZCD(Zero-Image Contrastive Decoding)는 대조적 디코딩(Contrastive Decoding)과 적응형 개연성 제약(Adaptive Plausibility Constraint)의 두 단계로 구성된다. 그림 1은 ZCD의 전체 구조를 나타낸다.

가. 대조적 디코딩

ZCD의 핵심 아이디어는 원본 이미지 v 와 0으로만 채워진 영 이미지 v_0 를 각각 모델에 입력하여 두 개의 출력 분포를 생성하고, 그 차이를 활용해 시각 정보를 더 강조하는 것이다. 영 이미지는 모든 픽셀 값이 0으로만 채워져 있어 추론에 필요한 시각 정보를 전혀 포함하지 않기 때문에 이에 기반한 출력 분포는 언어 편향에만 의존하게 된다. 이 두 분포의 차이를 이용해 VCD[2]의 식과 유사하게 새롭게 조정된 분포를 계산한다.

Setting	Method	Acc. ↑	F1 ↑
Random	Regular	0.829	0.808
	VCD	0.848	0.834
	ZCD	0.895	0.888
Popular	Regular	0.811	0.792
	VCD	0.825	0.813
	ZCD	0.873	0.868
Adversarial	Regular	0.786	0.771
	VCD	0.799	0.791
	ZCD	0.834	0.835

표 1 POPE 결과. 가장 좋은 성능을 볼드체로 표기하였음.

Method	Acc. ↑	Acc+ ↑	MME ↑
Regular	0.68	0.46	114.20
VCD	0.71	0.50	121.44
ZCD	0.77	0.59	136.31

표 2 MME 결과. Accuracy+는 한 이미지에 대한 두 질문을 모두 맞춘 결과의 비율이며, MME score는 Accuracy와 Accuracy+를 더한 점수임.

Method	Generative task			
	CHAIR ↓	Cover ↑	Hal ↓	Cog ↓
Regular	11.9	49.6	48.8	4.4
VCD	9.8	51.2	43.3	4.7
ZCD	8.0	53.3	40.6	4.3

Method	Discriminative task				
	Acc. ↑	Prec. ↑	Recall ↑	F1 ↑	AMBER ↑
Regular	74.7	81.1	80.7	80.9	84.5
VCD	75.9	82.1	81.3	81.7	86.0
ZCD	80.1	88.7	80.2	84.2	88.1

표 3 AMBER 결과. CHAIR는 총 객체 중 환각 객체의 비율이며 Cover는 총 객체 중 정답 객체의 비율임. Hal은 환각 객체가 포함된 문장의 비율, Cog는 인간의 인지와 비슷한 환각 객체가 포함된 비율임. AMBER score는 CHAIR와 F1 점수를 종합적으로 고려한 지표임.

$$p_{zcd}(y|v, v_0, x) = \text{softmax}[(1 + \alpha)\text{logit}_\theta(y|x, v) - \alpha \text{logit}_\theta(y|x, v_0)] \quad (2)$$

이때, α 는 두 출력 분포의 차이의 반영 정도를 결정하는 하이퍼파라미터(hyperparameter)로, $\alpha = 0$ 이면 일반적인 디코딩과 동일하게 작동한다.

나. 적응형 개연성 제약

영 이미지로부터 얻은 출력 분포는 왜곡된 시각 정보를 포함하고 있지만 기본적인 문법이나, 상식에 대한 정보 또한 여전히 포함하고 있다. 이러한 특성을 고려하지 않고 단순히 α 를 통해 해당 출력 분포를 감소시키는 것은 오히려 문법적으로 부자연스럽거나 의미가 부정확한 토큰을 선택하게 할 수 있다. 따라서 이러한 현상을 막기 위해 기존 연구[3]에서 제안된 방법과 유사하게 적응형 개연성 제약을 적용한다. 이 제약은 원본 시각 입력에 기반한 출력 분포에서 일정 수준 이상 신뢰할 수 있는 토큰 집합에서만 다음 토큰을 선택할 수 있도록 한다.

$$\begin{aligned} \mathcal{V}_{head}(y_{<t}) &= \{y_t \in \mathcal{V}: \\ p_\theta(y_t|v, x, y_{<t}) &\geq \beta \max_{\omega} p_\theta(\omega|v, x, y_{<t})\} \\ p_{zcd}(y_t|v, v_0, x) &= 0, \text{ if } y_t \notin \mathcal{V}_{head}(y_{<t}) \end{aligned} \quad (3)$$

이때 $\mathcal{V}$ 는 LVLM의 총 출력 어휘 집합이며, 토큰 후보 집합 $\mathcal{V}_{head}$ 는 원본 시각 입력에 기반한 출력 분포에서 확률이 $\beta \cdot \text{최대 확률}$ 이상인 토큰들로 정의된다. β 는 필터링의 강도를 결정하는 하이퍼파라미터이다. β 값이 커질수록 높은 확률인 토큰을 제외한 나머지 토큰을 다음 후보에서 제외하게 되므로 더 공격적인 필터링이 가능해진다. 두 기법에 기반한 ZCD의 최종 디코딩은 다음과 같다.

$$\begin{aligned} y_t &\sim \text{softmax}\{(1 + \alpha)\text{logit}_\theta(y|x, v) \\ &\quad - \alpha \text{logit}_\theta(y|x, v_0)\} \\ \text{subject to } y_t &\in \mathcal{V}_{head}(y_{<t}) \end{aligned} \quad (4)$$

IV. 실험

ZCD 기법의 성능을 검증하기 위해, 대표적인 LVLM인 LLaVA1.5 모델[6]에서 실험을 적용하였으며, 다음 세 가지 설정을 비교했다.: (1) Baseline인 기본 LLaVA1.5 (2) LLaVA1.5 + VCD (3) LLaVA1.5 + ZCD. 평가는 LVLM의 환각 평가에 자주 사용되는 다음 벤치마크를 활용했다.: (1) 객체 존재 여부에 관한 POPE[4], (2) 14개의 멀티모달 과제를 포함하는 MME[1], (3) 단답형, 생성형 벤치마크를 모두 포함하는 AMBER[7]. 디코딩 방식은 baseline과 VCD, ZCD 방식 모두 샘플링 기반으로 진행하였다. 하이퍼파라미터 세팅은 VCD의 경우 논문[2]의 세팅과 동일하게 $\alpha = 1.0, \beta = 0.1$ 로 설정하였으며, ZCD의 경우 성능이 가장 높았던 $\alpha = 1.0, \beta = 0.4$ 로 설정했다. 모든 실험은 seed=55에서 수행했다.

표 1과 표 2에서 확인할 수 있듯이, ZCD는 POPE와 MME와 같은 discriminative task에서 기존 baseline과 VCD 기법보다 더 우수한 성능을 보였다. POPE 정확도에서 Baseline 대비 세팅 별로 6.6%p, 6.2%p, 4.8%p, VCD 대비 4.7%p, 4.8%p, 3.5%p 향상을 보였다. MME score에선 baseline 대비 22.11 점, VCD 대비 14.87 점 향상하였다. 표 3에서의 결과와 같이 AMBER와 같은 생성형 벤치마크에서도 우수한 성능을 보여주었다. 생성형·단답형 과제 모두 반영한 AMBER score에서 baseline 대비 3.6 점, VCD 대비 2.1 점 향상하였다.

이러한 결과는 영 이미지 기반으로 부정 로짓을 생성하는 ZCD 기법이 기존 기법인 VCD보다 더 효과적으로 생성형·단답형 과제 모두 객체 환각을 제어할 수 있음을 보여준다.

V. 결론

본 논문에서는 거대 시각 언어 모델에서의 객체 환각을 완화하기 위해 이미지와 관련이 없는 영 이미지에 기반한 부정 로짓을 활용하는 ZCD 기법을 제안하였다. LLaVA1.5 기반 실험 결과 ZCD는 베이스라인, VCD 기법대비 POPE, MME, AMBER와 같은 다양한 벤치마크에서 일관적인 성능 향상을 보였다. 이는 ZCD가 학습이나 추가적인 모델 없이도 LVLM의 환각 완화에 강력하고 실용적인 방안이 될 수 있음을 보여준다. 향후 연구에서는 ZCD를 다양한 디코딩 전략이나 환각 제어 기법과 결합하여 언어 편향을 더 효과적으로 제어하여 LVLM의 성능을 더욱 강화할 수 있는 기법을 탐구하고자 한다.

ACKNOWLEDGMENT

이 논문은 2025 년도 정부(과학기술정보통신부)의 재원으로
2025 년 혁신거점 인공지능 데이터 융합과제 사업의 지원을
받아 수행된 연구임(S2201-24-1002)

This thesis was conducted with the support of the 2025
innovation base artificial intelligence data convergence
project with the funding of the 2025 government (Ministry of
Science and ICT) (S2201-24-1002)

참 고 문 헌

- [1] C. Fu, P. Chen, Y. Shen, Y. Qin, M. Zhang, X. Lin, J. Yang, X. Zheng, K. Li, X. Sun et al., "Mme: A comprehensive evaluation benchmark for multimodal large language models," arXiv:2306.13394, 2023
- [2] Leng, Sicong, et al. "Mitigating object hallucinations in large vision-language models through visual contrastive decoding." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024.
- [3] Li, Xiang Lisa, et al. "Contrastive Decoding: Open-ended Text Generation as Optimization." Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2023.
- [4] Li, Yifan, et al. "Evaluating Object Hallucination in Large Vision-Language Models." *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. 2023.
- [5] Liu, Hanchao, et al. "A survey on hallucination in large vision-language models." *arXiv preprint arXiv:2402.00253* (2024).
- [6] Liu, Haotian, et al. "Improved baselines with visual instruction tuning." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2024.
- [7] Wang, Junyang, et al. "Amber: An llm-free multi-dimensional benchmark for mllms hallucination evaluation." *arXiv preprint arXiv:2311.07397* (2023).
- [8] Yan, Hong, et al. "Overcoming language priors with self-contrastive learning for visual question answering." *Multimedia Tools and Applications* 82.11 (2023): 16343-16358.
- [9] Zhao, Jianfei, et al. "Cross-image contrastive decoding: Precise, lossless suppression of language priors in large vision-language models." *arXiv preprint arXiv:2505.10634* (2025).