세분화된 hallucination 완화를 위한 VLMs의 DPO 기반 정렬 학습 연구

이종호[1], 김지훈[2], 이지은[3], 신용태[4]
[1,2]숭실대학교 컴퓨터학과 석사과정
[3]숭실대학교 컴퓨터학과 박사과정
[4]숭실대학교 컴퓨터학부 교수
leejongho@Soongsil.ac.kr, jihunthank@Soongsil.ac.kr, lhsgsse10@Soongsil.ac.kr,
shin@ssu.ac.kr

A Study on DPO-Based Alignment for Fine-Grained Hallucination Mitigation in VLMs

Jong-Ho Lee[1], Ji-Hun Kim[2], Ji-Eun Lee[3], Yong-Tae Shin[3]
[1,2,3]Dept. of Computer Science and Engineering, Soongsil University
[4]School. of Computer Science and Engineering, Soongsil University

요 약

본 연구는 Vision-Language Model(VLM)의 세분화된 hallucination 오류를 완화하기 위한 Direct Preference Optimization 기반 정렬 학습 기법을 제안한다. 사전 학습된 VLM이 생성한 이미지 캡션을 GPT-4o를 활용하여 명사구 단위로 분해하고, 각 명사구의 객체 및 속성 정보를 추출한 뒤 시각적 정합성을 평가하여 선호-비선호 응답 쌍을 생성한다. 생성된 응답 쌍은 별도의 보상 모델 없이 DPO 학습에 활용된다. COCO 2014 데이터셋을 기반으로 한 실험 결과, CHAIR 및 FaithScore 지표 모두에서 기존 대비 향상된 성능을 보였으며, 객체 및 속성 수준의 hallucination 완화 효과를 보였다. 이를 통해 제안하는 정렬 학습 기법이 VLM의 신뢰성을 향상시키는 데 효과적임을 보여준다.

1. 서론

Vision-Language Model(VLM)은 이미지와 텍스트를 동시에 이해하는 멀티모달 모델로 최근 Image Captioning이나 Visual Question Answering 등 다양한 시각-언어 과제에서 뛰어난 성능을 보이고 있다. 특히 VLM에 Large Language Model(LLM)과 결합하거나 LLM의 언어 생성 능력을 통합함으로써 문장 생성과 복잡한 추론 능력이 크게 향상되어 다양한 downstream task에서의 활용 가능성이 높아지고 있다 [1].

그러나 이러한 성능 향상에도 불구하고, VLM은 여전히 hallucination 문제를 심각하게 겪고 있다. Hallucination은 이미지의 실제 내용과 일치하지 않은 설명을 생성하는 현상으로, 예를 들어 고양이 이미지에 대해 개에 대한 설명을 생성하는 경우가 이에 해당한다. 이러한 오류는 이미지-텍스트 간 의미 정렬 실패나, 모델이 학습한 언어 패턴에 과도하게 의존하는 특성에서 기인하며, 특히 장문 생성이나 복잡한 추론이 요구되는 응답에서 더욱 빈번하게 나타난다. 이러한 오류는 의료, 자율주행 등 신뢰성이 핵심적인 분야에서 VLM의 활용을 제약하는 주요 원인 중 하나이다.

기존에는 이러한 문제를 완화하기 위한 방법으로 보상 모델과 강화학습을 결합한 Reinforcement Learning with Human Feedback(RLHF)이 사용되어 왔다. 그러나 RLHF는 보상 모델 학습의 복잡성, 낮은 안정성, 높은 비용 등과 같은 한계가 지적되어 왔다. 이를 해결하기 위한 대안으로 최근에는 별도의 보상 모델 없이 선호 응답 데이터만으로 정렬 학습을 수행할 수 있는 Direct Preference Optimization(DPO) 기법이 주목받고 있다 [2].

하지만 기존의 DPO 연구는 주로 텍스트 기반 언어 모델에 집중되어 있으며, VLM에서 발생하는 hallucination을 세부 수준에서 체계적으로 분석하고 정렬하는 시도는 부족하다. 특히, 이미지와 캡션 간의 정합성 오류를 객체(Object) 및 속성(Attribute) 수준에서 식별하고 이를 기준으로 학습에 활용하는 접근은 거의 이루어지지 않았다.

이에 본 연구는 VLM의 세분화된 hallucination 오류 완화를 위한 DPO 기반 정렬 학습 프레임워크를 제안한다. 본 프레임워크는 사전 학습된 VLM이

생성한 이미지 캡션을 명사구 단위로 분해한 후, 객체 및 속성 정보를 추출하여 시각적 정합성을 평가하여 정합성과 불일치 기준으로 응답 쌍을 구성하고 DPO 학습에 활용한다. 이를 통해 별도의 보상 모델 없이도 세분화된 오류를 억제하는 정렬 학습이 가능하며 결과적으로 VLM의 신뢰성을 효과적으로 향상시킬 수 있다.

2. 관련 연구

2.1 DPO

DPO는 인간의 선호 응답 쌍을 기반으로 언어 모델을 정렬하는 효과적인 기법으로 제안되었다. 기존의 정렬 방식인 RLHF는 보상 모델 학습과 Proximal Policy Optimization와 같은 강화학습 알고리즘이 필요하여 구현이 복잡성과 계산 비용 측면에서 한계를 지닌다.

DPO는 이러한 문제를 해결하기 위해 별도의 보상 모델 없이, 입력 x에 대해 선호 응답 y_w가 비선호 응답 y_l보다 더 높은 상대적 likelihood를 갖도록 학습 중인 정책 π_θ를 직접 최적화한다. 학습은 다음과 같은 이진 분류 기반 손실 함수를 통해 수행된다.

$$\iota_{DPO}(\theta) = -\log\sigma(\beta \cdot (\log\frac{\pi_\theta(y_w|x)}{\pi_{ref}(y_w|x)} - \log\frac{\pi_\theta(\pi_l|x)}{\pi_{ref}(\pi_l|x)})) \quad (1)$$

수식 (1)에서 π_θ는 학습 중인 policy 모델, π_{ref}는 초기 reference 모델, σ는 sigmoid 함수이며, β는 최적화의 민감도를 조절하는 하이퍼파라미터이다. 이 손실 함수는 선호 응답과 비선호 응답 간의 log-likelihood 차이를 확률적 판단 기준으로 전환하여, 선호 응답의 확률을 높이는 방향으로 모델을 정렬시킨다. 이를 통해 DPO는 강화학습 없이도 직관적이며 효율적인 선호 기반 정렬 학습이 가능하다.

2.2 DPO 기반 VLM

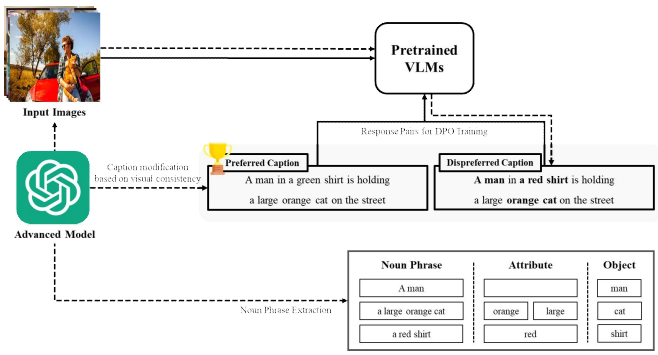
최근에는 DPO를 텍스트 기반 언어 모델뿐 아니라 VLM에도 적용하려는 연구가 진행되고 있다. 대표적인 연구로는 V-DPO[3]가 있으며, 이는 시각적 hallucination을 완화하는 정렬 학습 프레임워크로 제안되었다. 해당 연구는 이미지에 기반한 응답 쌍을 구성하고, 인간의 선호 판단을 통해 상대적으로 적절한 응답을 선택하여 DPO 방식으로 학습시킨다. 특히, V-DPO는 언어 기반 응답이 이미지 정보를 무시하는 현상을 방지하기 위해 Classifier-Free Guidance를 도입하고, 시각 정보 기반 likelihood를 보정함으로써 정렬 신호를 강화하였다. 이를 통해 기존의 RLHF 방식보다 효율적이고 안정적인 학습이 가능하며, 시각-언어 정합성이 개선되는 효과를 보였다.

그러나 V-DPO는 전체 응답을 단일 단위로 평가하고 정렬하는 구조로 설계되어 있어, 문장 내 특정 객체나 속성이 실제 이미지와 일치하지 않는 세부적인 오류를 식별하거나 이를 기준으로 학습하는 데 한계가 있다. 즉, 문장 전체의 정합성 여부만을 판단할 수 있어, 정밀한 오류 구분이나 객체와 속성 단위 학습이 어려운 구조이다.

본 연구는 이러한 한계를 극복하고자, 객체 및 속성 수준에서의 정합성 오류를 체계적으로 분석하고, 이를 기반으로 응답 쌍을 구성하여 DPO 학습을 수행하는 정렬 프레임워크를 제안한다. 이를 통해 VLM의 시각적 정합성과 세밀한 표현 신뢰성을 동시에 향상시키고자 한다.

3. DPO 기반 캡션 정합성 학습 파이프라인

본 연구에서는 VLM에서 발생하는 hallucination 오류를 객체 및 속성 수준에서 세분화하여 분석하고 정렬하기 위한 정렬 학습 파이프라인을 제안한다. 제안하는 프레임워크는 크게 두 단계, 즉 (1) 응답 쌍 생성과 (2) DPO 기반 학습으로 구성되며 전체 구조는 그림 1에 나타낸다.



(그림 1) 세분화된 hallucination 완화를 위한 정렬 학습 파이프라인 구조

3.1 명사구 정합성 기반 응답 쌍 생성

DPO 학습을 위한 데이터 구축을 위해, 먼저 사전 학습된 VLM으로 이미지의 캡션을 생성한다. 이후 생성된 캡션에서 Advanced Model(예: GPT-4o)을 이용해 전처리 과정을 거치며 다음과 같은 단계로 구성된다. 우선, 캡션에서 대명사나 'this image'와 같은 구체성이 떨어지는 명사구는 제외하고, 실

<표 1> 학습 데이터 구축을 위한 프롬프트 요약

목적	프롬프트 요약
정합성 검증	"Given the image, check if each noun phrase below correctly describes the image. Respond with 'correct' or 'incorrect' for each."
비선호 응답 생성	"Modify a given noun phrase so that it sounds natural but no longer matches the image."
선호 응답 생성	"Given a caption with an incorrect phrase, replace it with one that correctly matches the image and caption context.“

제 객체나 장면을 지칭하는 구체적인 명사구만을 추출한다. 추출된 각 명사구는 속성과 객체 구성 요소로 분리되고, 이미지 내에 존재하는지 시각적으로 검증한다. 명사구의 구성 요소가 이미지에서 확인될 경우 정합(correct), 그렇지 않을 경우 비정합(incorrect)으로 판단된다. 정합성 판단 결과에 따라 캡션은 두 가지 방식으로 응답 쌍을 구성한다.

첫째, 명사구가 모두 정합한 경우에는 원본 캡션을 선호(preferred) 응답으로 간주하고, 문맥을 유지한 상태에서 명사구를 이미지와 불일치하도록 변형하여 비선호(dispreferred) 응답을 생성한다.

둘째, 명사구 중 일부가 비정합한 경우에는 원본 캡션을 비선호 응답으로 간주하고, 해당 명사구만을 이미지와 문맥에 일치하도록 수정한 캡션을 선호 응답으로 생성한다.

이와 같은 학습 방식으로 각 이미지에 대해 하나의 응답 쌍을 생성한다. 이 과정은 학습에 앞서 DPO 훈련 데이터셋을 구축하기 위한 전처리 단계로 작동하며 그림 1의 점선으로 표시된 경로에 해당한다. 표 1은 정합성 판단 및 응답 변형에 활용된 주요 프롬프트 예시를 요약한 것이다.

3.2 DPO 기반 정렬 학습

앞 절에서 생성된 응답 쌍을 이용하여 본 연구는 DPO 기반의 정렬 학습을 수행한다. 해당 학습은 동일한 이미지에 대해 선호 응답이 비선호 응답보다 더 높은 likelihood를 갖도록 모델을 정렬한다. 이때 입력은 두 개의 텍스트 응답이며, 학습 과정은 수식 (1)에서 제시된 확률비 기반 손실 함수를 따르게 된다. 이를 통해 모델은 이미지와 정합성이 높은 표현을 우선적으로 생성하도록 정렬되며, 결과적으로 hallucination 오류 가능성을 억제하게 된다. 이 학습 단계는 그림 1에서 실선 경로로 표시되며, VLM의 성능 향상을 위한 핵심 절차에 해당한다.

4. 성능평가

4.1 실험 환경 및 데이터셋 구성

실험은 NVIDIA A100 GPU 환경에서 수행되었으며, LLaVA-1.5-7B 모델에 4-bit 양자화 및 LoRA 기반 fine-tuning을 적용하여 학습 효율을 높였다. 학습 데이터는 COCO 2014 validation 세트에서 무작위로 선택한 2000장의 이미지로 구성되며, 각 이미지에 대해 사전 학습된 LLaVA 모델로 생성한 캡션을 기반으로 데이터 구축을 진행하였다. 생성된 캡션으로부터는 GPT-4o를 활용해 구체적인 명사구를 추출하고, 각 명사구가 실제 이미지와 정합하는지를 시각 정보를 기반으로 정밀하게 판단하였다. 이러한 절차를 통해 각 이미지에 대해 정합성과 문맥을 고려한 선호 - 비선호 응답 쌍을 자동으로 구성하고, 이를 DPO 학습에 활용하였다. 이때 사용된 GPT-4o는 OpenAI에서 제공하는 API를 통해 호출하여 활용하였다.

4.2 평가 지표

모델의 캡션 생성 품질을 정량적으로 평가하기 위해, 두 가지 주요 hallucination 측정 지표인 CHAIR [4]와 FaithScore [5]를 활용하였다. CHAIR는 이미지 내에 존재하지 않는 객체의 언급을 기반으로 hallucination을 평가하며, 캡션에 포함된 전체 객체 중 잘못된 객체의 비율을 나타내는 $CHAIR_i$와, hallucination이 하나라도 포함된 문장의 비율을 나타내는 $CHAIR_s$의 두 가지 지표로 측정된다. 이는 캡션 전반의 사실성을 판단하는 데 유용하다.

FaithScore는 이미지와 캡션 간의 정합성을 평가하기 위한 모델 기반으로 정량화한 지표로, 이를 확장하여 객체와 속성 수준으로 별도 세분화하여 평가하였다. 각 명사구의 객체 및 속성이 실제 이미지와 시각적으로 정합하는지를 평가함으로써, 모델이 세부 정보에 대해 얼마나 신뢰할 수 있는지를 보다 정

밀하게 측정할 수 있다. 이 방식은 단순한 전체 문장 평가보다 세분화된 hallucination 완화 성능을 보다 구체적으로 나타낼 수 있도록 한다.

4.3 실험 결과

제안한 DPO 기반 학습 파이프라인의 효과를 검증하기 위해, 학습 전후 모델의 hallucination 오류를 CHAIR 및 FaithScore 지표를 통해 비교하였다. 표 2는 정렬 학습 전후의 성능 차이를 정량적으로 보여준다. 실험 결과, DPO 학습을 적용한 모델은 객체 및 문장 수준 모두에서 hallucination 발생 비율이 감소하였으며, 동시에 객체 및 속성 수준의 시각 정합성 점수도 향상되었다. 특히, 객체 수준의 hallucination 감소가 가장 두드러졌으며, 이는 명사구 기반 오류 분석 및 정렬 전략이 효과적이었음을 시사한다. 또한 속성 단위의 정밀한 오류 정렬을 통해, 모델의 세분화된 시각 정보 이해 능력이 개선된 것으로 확인되었다. 이와 같은 결과는 VLM의 응답 신뢰성을 높이기 위한 DPO 기반 정렬 학습의 유의미한 가능성을 보여준다.

<표 2> 성능평가 결과

Model	CHAIRi	CHAIRs	FaithScore_Obj	FaithSocre_Attr
Baseline (LLaVa-7B)	0.18	0.64	0.197	0.189
Ours	0.15	0.56	0.243	0.195

5. 결론 및 향후 연구

본 연구는 VVLM에서 발생하는 세분화된 hallucination 오류를 완화하기 위해 DPO 기반의 정렬 학습 프레임워크를 제안하였다. 사전 학습된 VLM이 생성한 이미지 캡션을 명사구 단위로 분해한 뒤, 객체 및 속성 정보를 추출하고 시각적 정합성을 평가하여 선호-비선호 응답 쌍을 구성하고, 이를 학습에 활용하는 방식이다. 이 과정은 별도의 보상 모델 없이도 정렬 학습을 가능하게 하며, 결과적으로 모델의 사실성과 신뢰성을 향상시킬 수 있도록 설계되었다.

COCO 2014 validation 세트를 기반으로 한 실험을 통해, 제안한 방법은 기존 LLaVA 모델 대비 CHAIR 및 FaithScore 지표 모두에서 의미 있는 성능 향상을 보였다. 특히 객체 단위 hallucination의 억제 효과가 뚜렷하게 나타났으며, 속성 수준의 정합성 향상 또한 모델의 세부 정보 처리 능력이 개선되었음을 시사한다. 이와 같은 결과는 VLM 응답의 신뢰성을 높이기 위한 DPO 기반 정렬 학습의 실질적인 가능성을 보여준다.

한편, 본 연구는 다음과 같은 한계를 가진다. 첫째, 정합성 판단 과정이 GPT-4o와 같은 자동화된 모델에 의존하고 있어, 일부 주관적인 오류 가능성이 존재한다. 둘째, 실험에 사용된 VLM 아키텍처가 LLaVA-1.5에 국한되어 있어, 다양한 구조의 모델들에 대해 본 프레임워크가 얼마나 효과적으로 확장될 수 있는지는 후속 연구를 통해 검증할 필요가 있다.

따라서 향후 연구에서는 MiniGPT-4, Kosmos-2, OpenFlamingo 등 다양한 VLM에 본 프레임워크를 적용하여 일반화 가능성을 평가하고, POPE, CLIPScore 등 새로운 정합성 지표를 도입하여 평가 체계를 보다 정밀화할 계획이다. 또한 정합성 기반 응답 쌍 생성 방식과 정렬 기법을 보다 정교하게 다듬어, 모델의 신뢰성과 학습 효율을 함께 높이는 방향으로 연구를 확장할 예정이다.

Acknowledgements

"본 연구는 과학기술정보통신부 및 정보통신기획평가원의 SW중심대학사업의 지원을 받아 수행되었음 (2024-0-00071)"

참고문헌

[1] C. Zhang et al., “Hallucination of Multimodal Large Language Models: A Survey” arXiv preprint, arXiv:2310.11570, 2023.

[2] Rafailov R., Sharma A. et al., “Direct preference optimization: your language model is secretly a reward model”, Proc. NeurIPS, New Orleans, 2023, pp. 53728 - 53741.

[3] Xie Y., Li G. et al., “V-DPO: Mitigating hallucination in large vision language models via vision-guided direct preference optimization”, Findings of ACL: EMNLP 2024, Miami, 2024, pp. 13258 - 13273.

[4] Rohrbach A., Hendricks L.A. et al., “Object hallucination in image captioning”, Proc. EMNLP 2018, Brussels, 2018, pp. 4035 - 4045.

[5] Jing L., Li R. et al., “FaithScore: Fine-grained evaluations of hallucinations in large vision-language models”, Findings of ACL: EMNLP 2024, Miami, 2024, pp. 5042 - 5063.