사실 기반 지식 증류와 DPO 기반 다단계 학습을 통한 조선 산업 현장 이미지 캡셔닝

전윤모 [1], 이승헌 [2], 김웅섭 [3]
[1] 동국대학교 정보통신공학과 석사과정
[2] 동국대학교 정보통신공학과 석사과정
[3] 동국대학교 정보통신공학과 교수

jumo0716@dongguk.edu[1], leesh914@dgu.ac.kr[2], woongsup@dongguk.edu[3]

Image Captioning for Shipyard Process Analysis via Grounded Truth-Aware Knowledge Distillation and Direct Preference Optimization

YoonMo Jeon, SeungHeon Lee, Woongsup Kim
Dept. of Information Communication Engineering, Dongguk University, Seoul, Republic of Korea

요 약

본 연구는 조선소 공정 자동화 및 안전 관리를 지원하기 위해, 산업 현장 이미지 캡셔닝 모델의 성능을 고도화하는 방법을 제안한다. 기존 범용 시각-언어 모델(VLM)은 복잡한 공정과 특수 장비를 정밀하게 서술하는 데 한계를 가진다. 이를 해결하기 위해 본 연구는 ① 교사 모델(LLaVA)과 객체 탐지 정보를 활용한 사실 기반 지식 증류(Grounded Truth-Aware Knowledge Distillation)를 통해 학생 모델(BLIP-2)에 도메인 지식과 표현 스타일을 이식하고, ② 이후 선호도 직접 최적화(Direct Preference Optimization, DPO)를 적용하여 캡션의 상세성과 정확성을 개선하는 다단계 학습 파이프라인을 제시한다. 조선소 현장에서 촬영한 5,000 장의 이미지 데이터셋으로 학습하고 별도의 800 장으로 평가한 결과, 제안 모델은 사전학습된 BLIP-2 대비 BLEU-4, CLIP Similarity 등 정량 지표에서 큰 향상을 보였으며, 정성적 평가에서도 전문가 수준에 가까운 상세 캡션을 생성함을 입증하였다. 특히 경량 모델 기반에서도 이러한 성능을 달성할 수 있음을 확인하여, 조선소 공정 기록, 안전 관리, 디지털 트윈 구축 등 산업 응용에 활용될 높은 잠재력을 제시한다.

1. 서론

최근 인공지능 기술의 발전에 힘입어 산업 현장의 디지털 전환이 가속화되고 있다. 특히 이미지와 언어를 동시에 처리하는 시각-언어 모델(VLM)은 CCTV, 드론 등으로 수집된 방대한 시각 데이터를 분석하여 공정을 기록하고 안전을 관리하는 데 핵심적인 역할을 할 것으로 기대된다. 조선소는 선박 건조를 위해 넓은 부지에서 다수의 중장비와 인력이 복잡하게 상호작용하는 대표적인 산업 현장으로, 시각적 상황을 정확히 이해하고 문서화하는 기술은 생산성과 안전성을 혁신할 잠재력을 지닌다.

하지만 현재의 고성능 VLM 들은 일반적인 상황 묘사에는 능숙하지만, “이동식 크레인이 선체 블록을 조립 위치로 옮기고 있다” 같이 조선소의 특수한 장비와 공정을 설명하는 데는 한계를 보인다. 이러한 한계는 최근 연구에서도 보고되고 있는데, 다양한 산업 및 의료 도메인에서 최신 VLM 들이 일반적 묘사에는 강점을 보이지만 특수 도메인에서는 성능이 저하된다고 평가하였으며 [1], 도메인 변동(domain shift)에 취약하다는 점을 실험적으로 규명하였다 [2].

본 연구는 이러한 한계를 극복하고자, 저자들이 이전 연구에서 수행했던 접근법을 기반으로 더욱 고도화된 방법론을 제안한다. 우리의 이전 연구 [3]에서는 대형 교사 모델의 지식을 소형 학생 모델로 전달하는 지식 증류(Knowledge Distillation) 기법 [4]을 통해, 엣지 디바이스에서의 동작을 목표로 하는 모델 경량화의 가능성을 성공적으로 입증하였다. 본 논문은 여기에서 한 단계 더 나아가, 모델의 질적 성능을 극한으로 끌어올리는 것에 초점을 맞춘다.

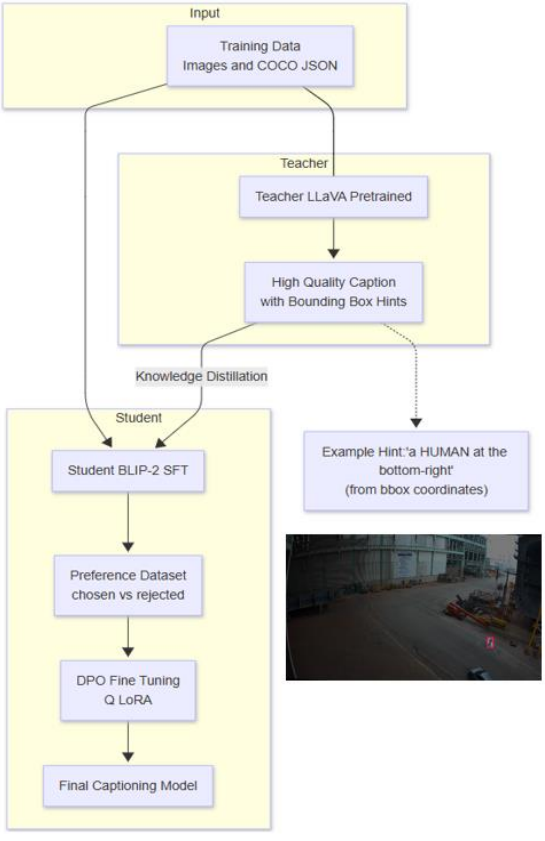
이를 위해 본 연구는 (그림 1)과 같이 다단계 학습

파이프라인을 설계하였다. 첫째, 단순 지식 증류를 넘어, 이미지 내 객체의 위치 정보(Bounding Box)를 명시적으로 활용하여 교사 모델(LLaVA)이 사실에 입각한 캡션을 생성하게 하는 '사실 기반 지식 증류(Grounded Truth-Aware Knowledge Distillation)'를 수행한다. 둘째, AI 피드백 기반 강화학습(RLHF)의 불안정성을 극복하기 위한 대안으로 안정성과 효율성이 입증된 '선호도 직접 최적화(DPO)' [5]를 도입하여, 모델이 더 상세하고 전문가가 선호하는 방향의 표현을 학습하도록 미세 조정한다. 이처럼 본 연구는 이전 연구의 성과를 계승하면서도, 근거 기반 데이터 생성과 DPO 라는 새로운 기법을 도입하여 도메인 특화 이미지 캡셔닝의 성능을 새로운 차원으로 끌어올리는 포괄적인 방법론을 제시한다는 점에서 차별점을 갖는다.

2. 본론

2.1 학습 파이프라인 모델

본 연구는 학생 모델로 효율적인 구조의 BLIP-2 [6]를, 교사 모델로 LLaVA 를 채택하였다. 조선소 환경을 담은 5,000 장의 이미지와 COCO JSON 형식의 어노테이션 데이터셋을 기반으로, (그림 1)과 같이 지식 증류와 선호도 튜닝의 다단계 파이프라인을 통해 모델을 학습시켰다.



(그림 1) 다단계 학습 파이프라인 모델

첫 단계인 지식 증류에서는 먼저 교사 모델(LLaVA)이 고품질의 학습 데이터를 생성한다. 이 과정의 핵심은 COCO JSON 데이터셋의 바운딩 박스 정보를 “이미지 오른쪽 하단에 HUMAN 객체가 있음”과 같은 근거 텍스트(Grounded Text)로 자동 변환하여 교사 모델의 프롬프트에 주입하는 것이다. 이는 교사 모델이 이미지에 존재하지 않는 객체를 언급하는 환각(hallucination) 현상을 억제하고, 실제 객체의 위치를 기반으로 사실에 입각한(factual) 상세 캡션을 생성하도록 유도한다. 이렇게 구축된 고품질의 (이미지, 생성 캡션) 쌍을 활용하여 학생 모델에 대한 지도 미세조정(SFT)을 수행함으로써, 교사 모델의 전문 어휘, 문장 구조, 상세 묘사 스타일을 효과적으로 이식한다. 이 과정에서 사용된 지식 증류 손실 함수는 다음과 같다.

$$\mathcal{L}_{KD} = -\sum_{t=1}^{T} P_{teacher}(y_t|x) \log P_{student}(y_t|x)$$

SFT 를 통해 스타일이 이식된 모델을 인간의 선호도에 더 가깝게 미세 조정하기 위해, 기존 PPO 기반 강화학습의 복잡성과 불안정성을 극복하는 대안으로 선호도 직접 최적화(DPO)를 도입하였다. DPO 학습을 위해 먼저, SFT 모델이 생성한 기본 캡션을 ‘비선호(rejected)’ 응답으로 간주하고, 다시 교사 모델에게 이미지와 rejected 캡션을 함께 제시하며 이를 개선의 기준으로 삼아 질적으로 더 우월한 설명을 생성하도록 유도하는 프롬프트를 통해 ‘선호(chosen)’ 응답을 생성한다. 이렇게 구축된 (image, chosen, rejected) 선호 쌍 데이터셋을 사용하여, 별도의 보상 모델 없이 아래의 손실 함수를 통해 직접적으로 정책을 최적화하는 DPO 학습을 수행한다.

$$\mathcal{L}_{DPO} = -\log \sigma\left(\beta\left(\log \pi_\theta(y^+|x) - \log \pi_\theta(y^-|x)\right)\right)$$

여기서 πθ 는 학생 모델의 확률 분포, y+는 선호 응답(chosen), y−는 비선호 응답(rejected), σ 는 시그모이드 함수, β 는 스케일링 파라미터를 의미한다. 이 과정을 통해 모델은 단순히 옳은 설명을 넘어, 두 캡션 사이의 미묘한 표현 차이를 학습하여 더 상세하고 자연스러운 문장을 생성하도록 최종적으로 개선된다.

모든 학습 과정에는 Q-LoRA 기법을 적용하여, 제한된 GPU 메모리 환경(24GB)에서도 효율적인 학습이 가능하도록 구성하였다.

2.2 평가지표 및 실험결과

제안하는 파이프라인의 효과를 단계별로 검증하기 위해, 학습에 사용하지 않은 800 장의 별도 테스트 이미지셋을 사용하여 정량 및 정성 평가를 수행하였다. 비교 대상은 사전 학습된 모델(Pre-trained Baseline), 제안 방식으로 SFT 를 마친 모델(Ours (SFT)), 그리고 DPO 까지 완료한 최종 모델(Ours (DPO))의 세 가지로 구성했다.

정량 평가는 생성된 캡션과 교사 모델(LLaVA)이 생성한 참조 캡션(reference caption) 간의 n-gram 정밀도를 측정하는 표준 지표인 BLEU-4 를 사용하였다. 즉, BLEU-4 점수는 학생 모델이 교사 모델의 서술을 얼마나 충실히 재현하는지를 보여주며, 모델별 전체 성능 비교 결과는 <표 1>과 같다.

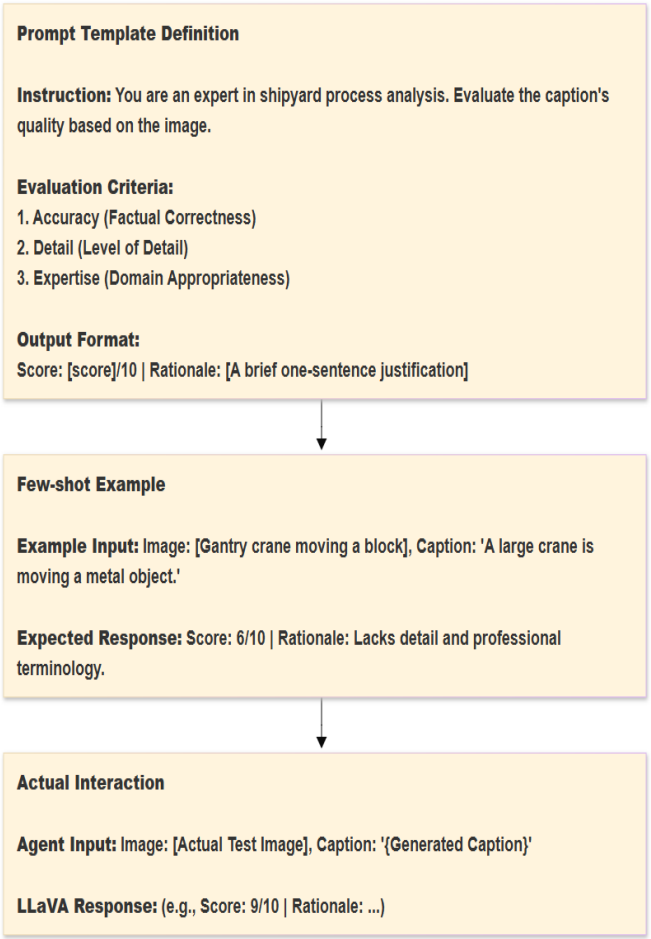
<표 1> 모델별 정량 및 정성 성능 비교

모델명	BLEU-4	CLIP Sim.(%)	LLaVA 평가 (평균)
Blip2(pre-train)	0.152	25.4	2.5 / 10
Ours (SFT)	0.351	31.5	8.5 / 10
Ours (DPO)	0.348	32.1	9.1 / 10

<표 1>에서 볼 수 있듯이, Baseline 모델 대비 SFT 를 적용하자 BLEU-4 점수가 0.152 → 0.351 로 크게 향상되었으며, 최종 DPO 모델은 BLEU-4 수치에서는 유사하지만 CLIP Similarity[7]와 LLaVA 평가에서 추가적인 성능 개선을 보여주었다.

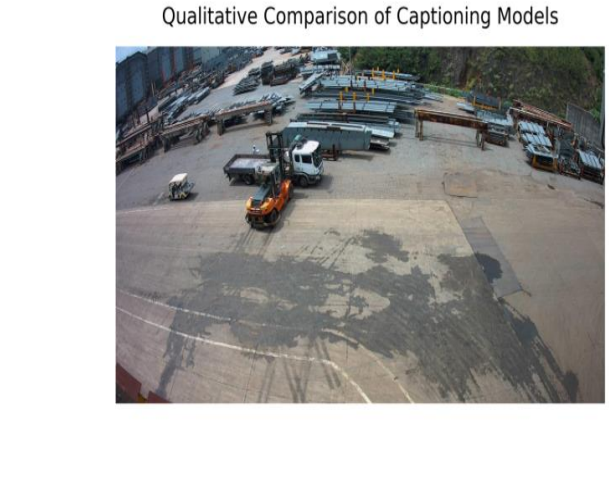
여기서 CLIP Similarity 는 이미지와 텍스트를 동일한 임베딩 공간에 매핑한 뒤, 두 벡터의 코사인 유사도를 계산하는 방식으로, 생성된 캡션이 실제 이미지 내용을 얼마나 잘 반영하는지를 정량적으로 측정하는 지표이다. 따라서 BLEU-4 가 낮게 나와도 CLIP 점수가 개선되었다면, 이는 모델이 단순히 단어를 맞추는 수준을 넘어 의미적 일관성을 확보했음을 의미한다.

그러나 BLEU 와 같은 전통적 정량 지표는 산업 현장에서 요구되는 전문적 표현을 충분히 반영하지 못한다. 따라서 본 연구는 이를 보완하기 위해 오픈소스 VLM 인 LLaVA 를 활용한 AI 기반 정성 평가를 도입하였다. (그림 2)는 평가에 사용된 프롬프트를 보여준다. LLaVA 가 '조선소 전문가'의 관점에서 캡션의 (1) 정확성, (2) 상세함, (3) 전문성을 종합적으로 판단하여 10 점 만점으로 채점하도록 설계했다. AI 기반 정성 평가 결과는 단계별 성능 향상을 가장 명확하게 보여준다. Pre-trained 모델은 2.5 점의 낮은 점수를 기록했으나, SFT 를 통해 8.5 점으로 도약했으며, 최종 DPO 모델은 9.1 점이라는 전문가 수준에 가까운 점수를 획득하였다.



(그림 2) LLaVA 정성 평가에 사용된 프롬프트 템플릿

(그림 3)은 제안한 학습 파이프라인이 생성한 캡션의 질적 차이를 시각적으로 보여준다. Pre-trained Baseline 모델은 "An industrial yard with a forklift, a truck, and stacks of metal beams"와 같이 단순 나열 수준의 일반적 묘사에 그쳤다. SFT 모델은 "In a shipyard's material storage area, an orange forklift is positioned near a white truck, while various steel sections are organized in the background"과 같이 구체적인 장면과 배치 정보를 포착하기 시작하였다. 반면, 최종 DPO 모델은 "A forklift and a flatbed truck are staged in a shipyard's block fabrication yard, prepared for material handling; in the background, numerous prefabricated metallic parts are methodically arranged, awaiting transport to the assembly area"와 같이 장면의 맥락과 목적까지 반영한 전문가 수준의 설명을 생성하였다. 이러한 결과는 제안된 다단계 학습 파이프라인이 단순 시각적 사실을 기술하는 데서 그치지 않고, 산업 현장에서 실제로 요구되는 맥락적 해석과 전문적 표현까지 효과적으로 학습했음을 입증한다.



(그림 3) 모델별 캡션 생성 결과 정성 비교

3. 결론

본 논문에서는 근거 기반 지식 증류와 선호도 직접 최적화(DPO)를 결합한 다단계 학습 파이프라인을 통해, 조선해양 분야 특화 이미지 캡셔닝 모델을 성공적으로 구현하였다. 제안 방법은 교사 모델과 객체 위치 정보를 활용해 학생 모델의 표현력을 강화하고, 이후 선호도 학습을 통해 품질을 점진적으로 고도화함으로써 PPO 기반 강화학습의 불안정성을 극복하는 효과적인 대안이 될 수 있음을 보였다. 실험 결과, BLEU-4 와 CLIP Similarity, AI 기반 정성 평가 모두에서 기존 대비 월등한 성능 향상을 확인하였다.

본 연구의 의의는 단순 객체 인식을 넘어, 산업 현장에서 요구되는 맥락적 이해와 전문가 수준의 서술을 가능하게 했다는 점이다. 특히 경량화된 BLIP-2 기반으로도 이러한 성능을 달성함으로써, 대규모 자원 없이도 실용적인 도메인 특화 VLM 을 구축할 수 있음을 입증하였다.

향후에는 조선해양 분야를 넘어 재난 대응, 건설, 의료 등 다양한 산업 도메인에 본 파이프라인을 확장 적용하여, 일반화 성능과 실용적 활용 가능성을 검증할 예정이다.

ACKNOWLEDGMENT

이 논문은 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원-학석사연계 ICT 핵심인재 양성 지원을 받아 수행한 연구임 (IITP-2024-00436744).

참고문헌

[1] Y. Jiang, X. Yan, G.-P. Ji, K. Fu, M. Sun, H. Xiong, D.-P. Fan, F. S. Khan, “Effectiveness Assessment of Recent Large Vision-Language Models,” Visual Intelligence, 2024.

[2] M. Koddenbrock, R. Hoffmann, D. Brodmann, and E. Rodner, “On the Domain Robustness of Contrastive Vision-Language Models,” arXiv preprint arXiv:2306.13663, 2023.

[3] 이승헌, 전윤모, 김웅섭, "BLIP 기반 Knowledge Distillation 을 활용한 멀티모달 이미지 캡셔닝 모델의 경량화 및 성능 개선 연구," 한국통신학회 하계종합학술발표회, 제주, 2025, pp. 1256-1257.

[4] G. Hinton, O. Vinyals, J. Dean, "Distilling the Knowledge in a Neural Network," arXiv preprint arXiv:1503.02531, 2015.

[5] R. Rafailov, A. Sharma, E. Mitchell, S. Ermon, C. D. Manning, and C. Finn, "Direct Preference Optimization: Your Language Model is Secretly a Reward Model," Advances in Neural Information Processing Systems (NeurIPS), Vol. 36**,** New Orleans, USA, 2023.

[6] J. Li, D. Li, S. Savarese, and S. Hoi, "BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models," 40th International Conference on Machine Learning (ICML), Honolulu, USA, 2023, pp. 19730-19742.

[7] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., "Learning Transferable Visual Models From Natural Language Supervision," 38th International Conference on Machine Learning (ICML), Virtual, 2021, pp. 8748-8763.